HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

MASTERARBEIT

Herr
Riccardo Brumm

Untersuchung des STAT3 und STAT1 Proteininteraktionsnetzwerks mittels Affinitätsmassenspektrometrie unter Einbeziehung komplementärer Daten

2012

Fakultät: Mathematik/Naturwissenschaften/Informatik

MASTERARBEIT

Untersuchung des STAT3 und STAT1 Proteininteraktionsnetzwerks mittels Affinitätsmassenspektrometrie unter Einbeziehung komplementärer Daten

Autor:
Riccardo Brumm

Studiengang:
Molekularbiologie/Bioinformatik

Seminargruppe:
MO10w1-M

Erstprüfer:
Dr. Stefan Kalkhof

Zweitprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Mittweida, 2012

Danksagung

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mich bei der Anfertigung dieser Arbeit unterstützt haben.

An erster Stelle gebührt mein Dank meinen beiden Betreuern Dr. Stefan Kalkhof und Prof. Dr. Dirk Labudde, die diese Arbeit ermöglichten und mich bei Bedarf mit wertvollen Ratschlägen unterstützt haben.

Bedanken möchte ich mich auch bei der Arbeitsgruppe von Dr. Stefan Kalkhof am UFZ, die bei Bedarf immer einige nützliche Ratschläge parat hatten. Dabei möchte ich mich besonders bei Jacqueline Kobelt für die Bearbeitung meiner Proben bedanken.

Weiterhin möchte ich mich bei der Arbeitsgruppe von Prof. Friedemann Horn von der Universität Leipzig, im Besonderen bei Gabriele Pfeiffer und Conny Blumert für die Anfertigung meiner Proben bedanken.

Ein Dank gebührt auch Bill Andreopoulos für die Prozessierung der Netzwerk-Motive sowie Dr. Michael R. Alvers und Matthias Zschunke, für die Unterstützung bei der Automatisierung der Anfragen an den Webservice von GoPubMed.

Weiterer Dank gebührt Dr. Martin von Bergen für die Möglichkeit meine Abschlussarbeit im Department Proteomik am Helmholtz-Zentrum für Umweltforschung – UFZ schreiben zu dürfen.

Höchste Wertschätzung gebührt all meinen Freunden und meiner Familie, die mir mit Ausdauer und Geduld in allen schwierigen Situationen der letzten fünf Jahre Rückhalt gegeben haben und mich abschließend eifrig im Kampf gegen Druckfehler und Stilblüten unterstützt haben.

Bibliografische Angaben

Brumm, Riccardo:
Untersuchung des STAT3/STAT1 Proteininteraktionsnetzwerks mittels Affinitätsmassenspektrometrie unter Einbeziehung komplementärer Daten.
81 Seiten, Hochschule Mittweida (FH), University of Applied Sciences,
Fakultät Mathematik/Naturwissenschaften/Informatik, Masterarbeit, 2012

Referat

Die Lokalisation, Aktivität, Funktion, Abbau sowie Synthese von Proteinen werden maßgeblich durch Wechselwirkungen von Proteinen mit weiteren Proteinen, anderen Biopolymeren sowie niedermolekularen Molekülen reguliert. Informationen über Art, Stärke und Abhängigkeit der Interaktionen sind daher von entscheidender Bedeutung für ein umfassendes Verständnis der Prozesse, in die ein Protein involviert ist, sowie den Mechanismen, durch die es reguliert wird. Die umfassende Charakterisierung von Interaktionen von Proteinen in einem gegebenen Proteom (mittlerweile oft als Interaktom bezeichnet) wird der nächste Meilenstein auf dem Weg zum Verstehen der Biochemie von den Zellen. Anormale Veränderungen von Protein-Protein- oder Protein-Metabolit-Interaktionen können Ursachen von Krankheiten sein, wohingegen gezielte medikamentöse Beeinflussungen Ansatzpunkte für Krankheitsbehandlungen darstellen.

Im Rahmen dieser Masterarbeit wurde das Interaktom der Proteine STAT3, STAT1, BMI1 und CDK9 in humanen embryonalen Nierenzellen (HEK 293T, engl. *Human Embryonic Kidney 293T cells*) mit einer auf Affinitätsmassenspektrometrie basierenden Strategie untersucht, in der stabile Isotopenmarkierung durch Aminosäuren in Zellkultur (SILAC), *in situ* Biotinylierung der vier ausgewählten Proteine, Affinitätsanreicherung und massenspektrometrische Analyse verbunden wurden. Den Schwerpunkt der Arbeit stellte die Optimierung der Datenauswertung dar. Zu diesem Zweck wurde eine Software entwickelt, die ein Protein-Protein Interaktionsnetzwerk aus Interaktionsdatenbanken um das jeweils zu untersuchende Protein erstellt und mit Hilfe von einer Meta-Datenbank und dem Protein-Protein Interaktionsnetzwerk die signifikanten Bindungspartner der Analyse selbstständig ermittelt. Die ermittelten Bindungspartner sollten mit Dreieck-Netzwerk-Motiven und komplementären Daten nachprozessiert bzw. nachevaluiert werden. Mit der PIPs Datenbank sollten alle Ergebnisse verglichen und evaluiert werden, weil diese Datenbank einen ähnlichen Ansatz mit komplementären Daten verfolgt und bereits seit einigen Jahren etabliert ist.

Investigation of the STAT3/STAT1 protein interaction network by affinity mass spectrometry including complementary data.

Abstract

The location, activity, function, degradation and synthesis of proteins are significantly regulated by interactions of proteins with other proteins, biopolymers and other low molecular weight molecules. Information about the type, strength and function of the interactions are therefore crucial for a comprehensive understanding of the processes, in which a protein is involved, and the mechanisms by which it is regulated. The comprehensive characterization of interactions of proteins in a given proteome (also known as interactome) is the next milestone on the path to understanding the biochemistry of the cell. Abnormal changes of protein-protein or protein-metabolite interactions may be the causes of disease, whereas targeted drug influences represent targets for disease treatments.

As part of this master's thesis the interactome of the proteins STAT3, STAT1, BMI1 and CDK9 was studied in human embryonic kidney 293T cells based on an affinity mass spectrometry strategy combined with stable isotope labeling by amino acids in cell culture (SILAC), *in situ* biotinylation of the four selected proteins, affinitypurification and mass spectrometry analysis. The focus of the thesis was the optimization of data analysis. Constituted for this purpose, a software was developed that creates a significant protein-protein interaction network of the obtained protein by using different interaction databases. This software determines binding partners independently by using that protein-protein interaction network supported by a meta-database. By using triangle network motifs and complementary data the identified binding partners ought to be judged and evaluated, respectively. The PIPs database was used for comparing and evaluation of the results, because this database is a similar approach with complementary data and also established for several years.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AD ..... activation domain
AP ..... affinity purification
BMI1 ..... Polycomb complex protein BMI-1
BRET ..... Bioluminescence Resonance Energy Transfer
CDK9 ..... Cyclin-dependent kinase 9
CMC ..... chronic mucocutaneous candidiasis
CoIP ..... co-imunoprecipitation
CSV ..... comma-separated value
DBD ..... DNA binding domain
ES ..... Evolutionäre Strategie
FRET ..... Fluorescence Resonance Energy Transfer
GUI ..... grafical user interface
HEK 293T ..... Human Embryonic Kidney 293T cells
IMEx ..... International Molekular Exchance
JAK ..... Januskinasen
MINT ..... Molecular INTeraction Datenbank
MIscore ..... Molecular interactions confidence Score
PcG ..... Polycomb group
PCR1 ..... Polycomb repressive complexes 1
PPI ..... Protein-Protein Interaktion, Protein-Protein Interaktionen
PPIN ..... Protein-Protein Interaktionsnetzwerk
PSI-MI ..... Proteomics Standard Initiative - Molecular Interactions
PSLT ..... Protein Subcellular Localization Tool
P-TEFb ..... positive transcription elongation factor b
PTL ..... posttranslational modification
SCOPPI ..... structural classification of protein-protein interfaces
SILAC ..... stable isotope labeling by amino acids in cell culture
STAT1 ..... Signal transducer and activator of transcription 1
STAT3 ..... Signal transducer and activator of transcription 3
UniProtKB-AN ..... UniProtKB accession number

Abbildungsverzeichnis

Tabellenverzeichnis

Vorwort

Die Lokalisation, Aktivität, Funktion, Abbau sowie Synthese von Proteinen werden maßgeblich durch Wechselwirkungen von Proteinen mit weiteren Proteinen, anderen Biopolymeren sowie niedermolekularen Molekülen reguliert. Informationen über Art, Stärke und Abhängigkeit der Interaktionen sind daher von entscheidender Bedeutung für ein umfassendes Verständnis der Prozesse, in die ein Protein involviert ist, sowie den Mechanismen, durch die es reguliert wird. Die umfassende Charakterisierung von Interaktionen von Proteinen in einem gegebenen Proteom (mittlerweile oft als Interaktom bezeichnet) wird der nächste Meilenstein auf dem Weg zum Verstehen der Biochemie von den Zellen. Anormale Veränderungen von Protein-Protein- oder Protein-Metabolit-Interaktionen können Ursachen von Krankheiten sein [Charbonnier et al., 2008].

Die rund 30.000 Gene des Humanen Genoms resultieren – durch verschiedene post-translationale Modifikationen und Splicing-Mechanismen – in schätzungsweise bis zu 10^6 Proteinen. Obwohl von einigen dieser Proteine erwartet werden kann, dass sie relativ isoliert ihren Funktionen nachgehen, agiert der Großteil voraussichtlich zusammen mit anderen Proteinen in Komplexen oder Netzwerken, um die Vielzahl an Prozessen zu organisieren, die Auswirkungen auf die zelluläre Struktur und die Funktion haben. Diese Prozesse beinhaltet Zellzyklus-Kontrolle, Zelldifferenzierung, Proteinfaltung, Signalübertragung, Transkription, Translation, Transport und posttranslationale Modifikationen. [Stelzl & Wanker, 2006]

Schlussfolgerungen über die Funktion von Proteinen können durch Studien über Protein-Protein Interaktionen (PPI) aufgestellt werden. Diese Schlussfolgerungen basieren darauf, dass die Funktion von unbekannten Proteinen durch die Interaktion mit einem Protein mit bekannter Funktion ermittelt werden kann. Grundlegend können PPIs stabil und langlebig oder weniger stabil und kurzlebig auftreten, wobei beide Arten wiederum entweder stark oder schwach ausgeprägt sein können. Zu den stabilen Interaktionen zählen Komplexe mit vielen Untereinheiten. Zwei Beispiele stabiler Komplexe mit mehreren Untereinheiten sind Hämoglobin und RNA-Polymerase. Stabile Interaktionen können durch Co-Immunpräzipitation, Pulldown oder far-Western blotting erforscht werden. Kurzlebige PPIs bestimmen voraussichtlich einen Großteil der zellulären Prozesse. Diese Interaktionen sind temporär in der Natur. Charakteristisch für deren Entstehung sind bestimmte Bedingungen in deren Umfeld. Diese kurzlebigen Interaktionen können stark und schwach, schnell und langsam sein. Vermutlich sind sie an der ganzen Bandbreite zellulärer Prozesse beteiligt, wenn sie in Kontakt zu ihren Bindungspartnern stehen. Verlässliche und umfassende Untersuchungen kurzlebiger sowie instabiler Interaktionen stellen jedoch noch immer eine große Herausforderung dar. **[**Brown & Jurisica, 2007; Boulon et al., 2010**]**

1 Grundlagen

1.1 Techniken zur Bestimmung von PPIs

Es gibt eine Reihe an Methoden, die der Bestimmung von Protein-Protein Interaktionen dienen. Die geläufigsten sind im Folgenden aufgeführt.

Das Mammalia Two-hybrid System basiert darauf, dass eukaryotische Transkriptionsfaktoren aus zwei unterschiedlichen Domänen bestehen: einer Domäne, die an die DNA bindet (DBD, engl. *DNA binding domain*) und einer für die Aktivierung zuständigen Domäne (AD, engl. *activation domain*). Das System beruht auf drei Plasmiden, die einer Säugetierzelle zugegeben werden. Ein Plasmid enthält das Bäuteprotein (bezeichnet als Y), das mit einer AD versehen ist (AD-Y). Ein zweites Plasmid enthält das Köderprotein (bezeichnet als X), das mit einer DBD versehen ist. Das dritte Plasmid ist mit der DNA Bindestelle für die DBD und einem Reportergen versehen. Kommt es zu einer Interaktion zwischen Protein X und Y wird die DBD durch die AD aktiviert und das Reportergen wird abgelesen. [Fearon et al., 1992]

Das Pulldown Assay ist eine weitere Möglichkeit, die auf Wechselwirkungen zwischen einem Fusionsprotein (das zu untersuchende Protein) und einem potentiellem Interaktionspartner basiert. Das Fusionsprotein wird dafür exprimiert und immobilisiert, wobei ein Ligand, der spezifisch für das Fusion Tag ist, angebracht wird. Das immobilisierte Fusionsprotein wird danach mit dem potentiellen Interaktionspartnern inkubiert, wobei die Quelle der potentiellen Interaktionspartner davon abhängt, ob neue Interaktionspartner gesucht oder alte Interaktionspartner bestätigt werden sollen. Nach einer Reihe von Waschschritten kann der gesamte Komplex herausgelöst und die Interaktionspartner bestimmt werden. [Vermeulen et al., 2006]

Eine weitere Methode ist die Co-Immunpräzipitation, bei der unter Verwendung eines Ganzzellextraktes Interaktionen bestätigt werden. In diesem Ganzzellextrakt liegen die Proteine in ihrer nativen Konformation und einem komplexen Gemisch von zellulären Komponenten vor, das für erfolgreiche Wechselwirkungen erforderlich sein kann. Zusätzlich sind in eukaryotischen Zellen die posttranslationalen Modifikationen gegeben, die in prokaryotischen Zellen nicht auftreten würden. Nachdem mittels Zelllyse das Zellextrakt unter nicht denaturierenden Bedingungen erhalten wurde, wird ein Antikörper, der spezifisch an das zu untersuchende Protein bindet, zugegeben. Mit diesem Antikörper wird der Proteinkomplex (das zu untersuchende Protein und der gebundene potentielle Interaktionspartner) immobilisiert und nach einer Reihe von Waschschritten kann der gesamte Komplex herausgelöst und die Interaktionspartner bestimmt werden. [Phizicky & Fields, 1995]

Kolokalisierungstechniken, wie FRET (Fluorescence Resonance Energy Transfer) und BRET (Bioluminescence Resonance Energy Transfer), benötigen keine vorbereitenden Schritte, wie die Co-Immunpräzipitation oder das Pulldown Assay zur Zellextraktgewinnung, womit die physiologischen Bedingungen gewahrt werden. FRET kann *in vivo* zur Interaktionsbestimmung zwischen Proteinen genutzt werden. Dabei werden zwei Proteine, zwischen denen eine Interaktion vermutet wird, mit verschiedenen Fluoreszensmolekülen – einem Donor und einem Akzeptor Fluorophor – markiert. Ist der Abstand zwischen beiden Fluoreszensmolekülen, aufgrund einer Interaktion zwischen beiden Proteinen, gering genug, findet nach Anregung des Donormoleküls eine Lichtemission vom Akzeptormolekül statt. [Gordon et al., 1998] Bei der BRET wird ein biolumineszierender Donor genutzt, womit sich die Anregung von Außerhalb erübrigt. Dadurch kommt es zu weniger Hintergrundrauschen, ausgelöst durch Autofluoreszen. [Xu et al., 1999]

1.1.1 Streptavidin / biotin Tag

Das Streptavidin-Biotin System ist eines der stärksten bekannten nicht kovalenten biologischen Interaktion ($K_d = 10^{-15}$ M) [Holmberg et al., 2005]. Schon 1985 wurde gezeigt, dass ein kleines Akzeptorpeptid *in situ* biotinyliert werden kann, wenn es an ein zu untersuchendes Protein gebunden ist und das bakterielle Protein BirA koexprimiert ist [Howard et al., 1985]. Dieses bio-Tag wurde durch ein Screening einer synthetischen Peptidbibliothek für BirA vermittelte Biotinylierung identifiziert [Schatz, 1993]. In diesem Prozess wurde eine optimale Sequenzlänge von 23 Aminosäuren für die *in situ* Biotinylierung ermittelt. In Säugetierzellen führt die Koexpression von mit dem bio-Tag versehenen Proteinen und BirA zu effizienter *in situ* Biotinylierung des markierten Proteins. Durch die geringe Größe des Tags erfährt das Protein keine Veränderungen hinsichtlich physikochemischer Eigenschaften, wie Diffusionsrate, Stabilität oder Hydrophobizität. Zusätzlich werden die Bindetaschen des Proteins durch die geringe Größe nicht beeinflusst. Wie in [Boer et al., 2003] beschrieben, kann die Biotinylierung genutzt werden, um markierte Proteine von kultivierten Säugetierzellen oder von transgenen Tieren in einem Schritt aufzureinigen.

1.1.2 Triple-SILAC

In dieser Arbeit wurde die stabile Isotopenmarkierung durch Aminosäuren in Zellkultur (SILAC, engl. *stable isotope labeling by amino acids in cell culture*) Methode genutzt mit der eine bestmögliche Vergleichbarkeit von Quantifizierungsdaten erreicht werden kann. Diese Methode nutzt den Ansatz, dass Säugetierzellen nicht alle Aminosäuren synthetisieren können und diese essentiellen Aminosäuren zu den Zellkulturen hinzugegeben werden müssen, um deren Wachstum zu unterstützen. Die essentiellen Aminosäuren werden isotopenmarkiert, zu den Zellkulturmedien

hinzugefügt und somit *in sito* oder auch *in vivo* [Krüger et al., 2008] in jedes neu synthetisierte Protein eingebaut. Nach einer bestimmten Anzahl an Zellteilungen wurde nahezu jedes Exemplar der bestimmten Aminosäure durch ihr isotopenmarkiertes Pendant ersetzt. Isotopenmarkierte Aminosäuren sind kommerziell erhältlich, wodurch die Anwendung erheblich erleichtert wird. Das Zellwachstum von Zellen mit derart modifizierten Aminosäuren ist unverändert im Vergleich zu Zellen mit unmodifizierten Aminosäuren, dokumentiert durch gleiche Zellmorphologie, gleiche Zellteilungszeit und unveränderter Fähigkeit zur Zelldifferenzierung. Die Zellen mit den markierten Aminosäuren können dann der entsprechenden Behandlung ausgesetzt werden, die untersucht werden soll. Unveränderte Zellen dienen als Referenz. Die Proteinpopulationen von beiden Ansätzen können anschließend geerntet und miteinander vermengt werden, weil die Markierung direkt in der Proteinsequenz eingebaut ist. Die aufgereinigten Proteine bzw. Peptide behalten das genaue Verhältnis der markierten zu den unmarkierten Proteinen bzw. Peptiden bei, wenn keine Proteinsynthese mehr stattfindet. [Ong et al., 2002] Der Triple-SILAC ist ein Ansatz, bei dem neben den unmarkierten Aminosäuren zwei verschieden markierte Aminosäuren eingesetzt werden, wobei deren Masse unterschiedlich ist. Somit entstehen Proteine in denen entweder die leichten, die mittelschweren oder die schweren Aminosäuren eingebaut werden. Dadurch ist es möglich neben einer Negativkontrolle, die Zellen auf zwei unterschiedliche Arten zu stimulieren, wobei alles simultan mittels Affinitätsmassenspektrometrie quantifiziert werden kann, wodurch eine bestmögliche Vergleichbarkeit der Quantifizierungsdaten ermöglicht wird. [Geiger et al., 2011]

Wird im Folgenden von Verhältnis (engl. *ratio*) gesprochen, ist damit, falls nicht genauer beschrieben, das Verhältnis des jeweiligen Ansatzes (H/L, M/L, H/M) gemeint.

1.1.3 Nachteile von Hochdurchsatzstudien

Protein-Protein Interaktionen werden typischerweise im kleinen Maßstab im Pulldown Verfahren oder Ähnlichen Techniken identifiziert. Diese haben jedoch den Nachteil, dass sie langsam und teuer sind, wodurch kaum ein umfassendes Bild des menschlichen Interaktoms erzielt werden kann. [Stelzl & Wanker, 2006]

Von Protein-Protein Interaktionsnetzwerken (PPINs), die mittels Hochdurchsatzstudien erstellt wurden, ist bekannt, dass sie viele Fehler beinhalten [Chiang et al, 2007; Yip & Gerstein, 2009]. Nicht alle Interaktionen treten zur selben Zeit und am selben Ort in den unterschiedlichen Zellulären Stadien auf [Zhang et al., 2008]. Das impliziert, dass ein PPIN, das aus einer Menge an binären Protein-Protein Interaktionen aufgebaut ist, oft unvollständig ist. So ist es nicht ungewöhnlich, dass Daten unterschiedlicher Studien geringe Überschneidungen aufweisen. Beispielsweise teilen zwei umfangreiche Untersuchungen des Humanen Interaktoms [Rual et al., 2005; Stelzl et al., 2005], welche jeweils einige tausend Interaktionen aufweisen, nur sechs gleiche Interaktionen [Deane et

al., 2002; Hoffmann & Valencia, 2003; Hart et al., 2006]. Es wird geschätzt, dass in einigen PPINs mehr als 50 % der Interaktionen falsch positive Interaktionen sind [Aloy, 2007; Scott & Barton, 2007] und zudem in einigen PPINs die geschätzte Rate an falsch negativen Interaktionen annähernd bei 90 % liegt [Sprinzak et al., 2003; D'haeseleer & Church, 2004; Stumpf et al., 2008].

1.1.4 Lösungsansätze

Ko-Immunpräzipitation (CoIP, engl. *co-imunoprecipitation*) und Affinitätsaufreinigung (AP, engl. *affinity purification*) tendieren zu einem hohen Betrag an Hintergrundrauschen durch unspezifisch gebundene Proteine [Boulon et al., 2010]. Die analytische Herausforderung ist es, eine umfassende Liste von potentiellen Interaktionspartnern mit wenigen falsch positiven Treffern zu bieten. Dies kann auf mehreren Wegen erreicht werden:

Die erste Möglichkeit besteht darin, ein stringentes Aufreinigungsverfahren mit mehreren Tags zu entwickeln [Rigaut et al., 1999], um die unspezifisch bindenden Proteine bei der Fällung zu reduzieren. Das verstärkte Waschen in dieser Strategie birgt jedoch das Risiko des Verlustes weniger stark gebundener Interaktionspartner.

Eine zweite Möglichkeit besteht darin, Interaktome zu vergleichen. Diese Interaktome werden unter Verwendung einer hohen Anzahl von verschiedenen Köder-Proteinen aufgestellt, wobei Proteine, die von mehreren Ködern ausgefällt werden als falsch positiv deklariert und verworfen werden. [Trinkle-Mulcahy et al., 2008]

Eine dritte Möglichkeit besteht darin, die spezifischen Bindungspartner durch die Quantifizierung der relativen Proteinmengen zu identifizieren, entweder bestimmt durch Köder-Proteine oder Pulldown Assays als Kontrolle. [Ranish et al., 2007, Vermeulen et al., 2008]

Die vierte Möglichkeit besteht darin, die gemessenen Interaktionspartner mit Hilfe von bioinformatischen Werkzeugen und Datenbanken nachträglich zu evaluieren. Dieser Ansatz wurde in dieser Arbeit angewandt, wobei aus Interaktionsdatenbanken ein PPIN erstellt und durch dieses PPIN die potentiellen Interaktionen mit Dreieck-Netzwerk-Motiven und komplementären Daten nachprozessiert bzw. nachevaluiert werden. [Andreopoulos et al., 2007a]

Ein Dreieck-Netzwerk-Motiv umfasst dabei zwei Proteine, die jeweils den gleichen Interaktionspartner haben, der damit für beide einen Nachbar zweiter Stufe darstellt. Beide Proteine selbst sind über komplementäres Wissen, wie strukturelle Daten, miteinander verknüpft. Somit entsteht ein PPI-CD-PPI Motiv, wobei CD eine Verbindung durch komplementäre Daten beschreibt. Dreieck-Netzwerk-Motive bilden die Grundbausteine von PPINs [Milo et al., 2002; Albert & Albert, 2004; Jin et al., 2007]. Sie lassen sich mit der Beobachtung begründen, dass Protein mit gemeinsamen Interaktionspartnern einer hohen Wahrscheinlichkeit nach die gleichen Funktionen aufweisen [Okada et al., 2005; Li et al., 2006; Andreopoulos et al., 2007a]. Nachbarn der zweiten Stufe in

PPINs sind funktionell ähnlich und können für funktionelle Vorhersagen genutzt werden. [Zhang et al., 2006; Chua et al., 2007; Chua et al., 2008]. Es wurde gezeigt, dass mit Dreieck-Netzwerk-Motiven die falsch positiven und falsch negativen Interaktionen in einem PPIN besser entdeckt werden können [Andreopoulos et al., 2009].

1.2 Genetische Algorithmen und Evolutionäre Strategien

Aus Sicht eines Ingenieurs stellt die Evolution ein sehr ausgefallenes und interessantes Optimierungsverfahren dar. Dieser Prozess begann als sich auf der Erde vor etwa 3,8 Milliarden Jahren das erste Leben entwickelte [Oberbeck & Fogleman, 1989]. In diesem Zeitraum sind zum Teil sehr komplexe Lebensformen mit erstaunlichsten Anpassungen entstanden. Charles Darwin gelang es zuerst diese Theorie ausführlich darzulegen. Im Wesentlichen beruht Darwins Evolutionstheorie auf drei Beobachtungen und Grundannahmen. Er beobachtete zum Einen die Tatsache, dass fast alle Lebewesen mehr Nachkommen hervorbringen, als überleben können. Weiterhin stellte er fest, dass zwischen Lebewesen einer Art sehr starke Ähnlichkeit auftreten kann, sie aber nie vollkommen identisch sind. Die dritte Grundannahme ist, dass sich nur die erblichen Varianten in der Folgegeneration durchsetzen, die sich im Kampf um Anpassung und Überleben bewährt haben. Im Laufe der Generationen führt die letzte Annahme zu einer Optimierung der Lebewesen einer Art. [Darwin, 1859]

Die Evolution ist somit vergleichbar mit einem Suchprozess. Dabei stellen die genetischen Informationen bzw. die Kombination aller möglichen Nukleotidbasen, die im menschlichen Chromosom vorkommen können, den Suchraum dar. Ziel dieses Suchprozesses ist das Finden von Erbanlagen, die dem Individuum die besten Chancen im Kampf ums Überleben bietet. Im Wesentlichen beruht die Evolution auf drei Prinzipien: der Mutation, der Rekombination und der Selektion. Deren Kombination ist eine geschickte Verbindung aus ungerichteten und gerichteten Suchprozessen. Durch Mutationen von Genen werden lediglich kleine Veränderungen im Erbgut hervorgerufen. Sinn und Zweck der Mutation liegt optimierungstechnisch darin begründet, dass lokale Optima überwunden werden können. Somit wird verhindert, dass sich eine Population auf ein lokales Optimum fixiert. Durch das zufällige Auftreten von Mutationen findet hier ein ungerichteter Suchprozess statt. Die Rekombination beschreibt die Neukombination von einzelnen Genen oder ganzen Gensequenzen der Eltern, um ein neues Kind zu beschreiben. In der Regel werden dabei funktional verbundene Gene seltener voneinander getrennt, als funktional entfernte Gensequenzen. Somit läuft die Rekombination teilweise nach statistischen Gesetzmäßigkeiten, teilweise aber auch rein zufällig ab. Daraus folgt, dass sie zwischen einem gerichteten und einem ungerichteten Suchprozess einzuordnen ist. Durch Selektion wird bestimmt, welche Ausprägungen

sich in der nächsten Generation verstärken und welche nicht. Demzufolge ist die Selektion ein gerichteter Suchprozess. Durch sich ändernde Umwelteinflüsse führt die Evolution ein ständiges Rennen gegen die Zeit. Arten, die sich nicht schnell genug anpassen, sterben aus. Demnach haben die am besten angepassten Individuen die größte Überlebenschance. Daher kommt der Ausdruck: „*Survival of the fittest*“. [Rechenberg, 1973; Goldberg, 1989; Holland, 1992; Nissen, 1997]

Es gibt zahlreiche weitere Variationen und Modifikationen dieser Algorithmen. Im Folgenden wurde nur auf die Details eingegangen, die in dieser Arbeit auch umgesetzt wurden.

1.2.1 Evolutionäre Strategie

Die relevante Erbinformation eines Individuums ist in reellen Zahlen codierbar. Ein Individuum wird somit durch einen Vektor reeller Zahlen und eine Population durch eine Menge derartiger Vektoren dargestellt. Dieser an Realzahlen orientierte Ansatz hat historische Gründe. Ingo Rechenberg, der die Basis der Evolutionsstrategien (ES) entworfen hat, beschäftigte sich viel mit Parameteroptimierung, dessen Lösungen in der Regel aus reellen Zahlen bestanden. [Schöneburg, 1994]

Die einfachste Form ist die (1+1)-ES. Dabei wird von einem Ur-Individuum – dem Elter – ausgegangen, dass genau ein Kind bzw. Nachkomme erzeugt. Hierfür wird von dem Elter eine Kopie erzeugt, die im Anschluss stochastisch verändert wird. Diese Duplikation des Elters entspricht dabei dem biologischen Prozess der identischen Replikation der DNS und die stochastische Modifikation entspricht dem biologischen Prozess der Mutation. Durch eine Qualitäts-funktion wird nun jedem Individuum ein Fitnesswert zugewiesen. Anschließend tritt die Selektion in Kraft und das bessere der beiden Individuen „überlebt“. Grundsätzlich besteht dieser Algorithmus aus drei Schritten: Duplikation, Mutation und Selektion. [Rechenberg, 1973]

Die (μ+λ)-ES ist eine Verallgemeinerung der (1+1)-ES. Es werden bei dieser Strategie aus μ Eltern genau λ Kinder erzeugt. Als Einschränkung gilt hierbei: $\lambda \geq \mu \geq 1$. Zur Erzeugung der λ Nachkommen werden aus μ Eltern λ Nachkommen ausgewählt. Die Auswahlwahrscheinlichkeit ist dabei gleich, sodass es notwendiger weise, bei $\lambda > \mu$, auch zur Mehrfachauswahl kommt. Im Extremfall ist so auch die λ-fache Auswahl eines Elter möglich. Nach der Duplikation werden die Kopien mutiert und deren Fitness berechnet. Wie bei der (μ+λ)-ES werden aus den λ Nachkommen und den μ Eltern die besten μ Individuen ausgewählt. Somit bleibt die Anzahl an Eltern in jeder Generation gleich. [Rechenberg, 1973]

Diesen Formalismus ergänzte Rechenberg, um nicht nur einzelne Individuen, sondern auch ganze Populationen simulieren zu können. So beschreibt die [μ‘+λ‘(μ+λ)]-ES, dass aus μ‘ existierenden Populationen λ‘ neue Populationen gebildet werden und aus allen (sowohl den existierenden, als auch den neuen) Populationen die besten ausgewählt werden. Jede λ‘ Population besteht dabei, wie bereits im vorherigen Abschritt beschrieben, aus μ Eltern, die λ Kinder erzeugen. Aus allen

Eltern und Kindern einer Population werden die besten μ Individuen ausgewählt. Zusätzlich kann das Konzept der Isolation eingeführt werden. Bei der $[\mu'+\lambda'(\mu+\lambda)^n]^m$-ES ist jede Population für *n* Generationen einer unabhängigen Entwicklung ausgesetzt, dass der isolierten Entwicklung getrennter Populationen in der Natur nachempfunden ist. Im Anschluss werden die besten μ' Populationen ausgewählt. Eine Beurteilung der Populationen, um sie miteinander vergleichen zu können, ist auf verschiedene Weise möglich. So kann die Fitness einer Population beispielsweise aus dem Durchschnitt aller Fitnesswerte der enthaltenen Individuen berechnet werden oder der beste Fitnesswert aller Individuen einer Population repräsentiert die Fitness einer Population. Eine derartige Simulation bietet sich besonders für parallele Rechnerstrukturen an, weil jede Population getrennt voneinander behandelt werden kann. Zum Terminieren des Algorithmus gibt es verschiedene Möglichkeiten. Die simpelste Möglichkeit ist es, die Anzahl an Gesamtdurchläufen *m* zu begrenzen. Ein anderer Ansatz zum terminieren des Algorithmus wäre ein Abbruch, wenn sich die Individuen der besten Population nicht mehr verändern. In dieser Problemstellung kann es aber passieren, dass sich bei geringfügiger Modifikation der Parameter die Fitnesswerte nicht ändern. Dies hat zur Folge, dass abwechselnd immer die gleichen Individuen gefunden werden und der Algorithmus nicht terminiert. Ein Abbruch anhand sich nicht verändernder Fitnesswerte ist auch nicht möglich, weil theoretisch mit unterschiedlichen Parametersätzen die gleichen Fitnesswerte erreicht werden kann. Somit würde der Algorithmus auch hier nicht terminieren. [Rechenberg, 1973; Nissen 1997]

1.2.2 Genetischer Algorithmus

Als Initialisierungsschritt des Algorithmus muss die Menge an möglichen Lösungen, der Lösungsraum, bestimmt werden. Eine einzelne Lösung in diesem Lösungsraum stellt hierbei ein Individuum und eine Lösungsmenge stellt eine Population dar. Ein Individuum ist somit ein Vektor aus allen Variablen, die es zu optimieren gilt. Die Startwerte dieser Variablen werden bei der Initialisierung für die späteren Anwendungen von genetischen Operatoren aus der reellen Codierung in eine binäre Form gebracht. Daraus ergibt sich, dass ein Individuum aus einem Vektor aus Bitstrings[1] besteht, die einen bestimmten Punkt im Lösungsraum repräsentieren. Diese ist in Abbildung 1 verdeutlicht.

[1] Ein Bitstring ist eine Zeichenkette aus Nullen und Einsen, die von rechts nach links die 2er Potenzen, beginnend bei 2^0, aufzeigen.

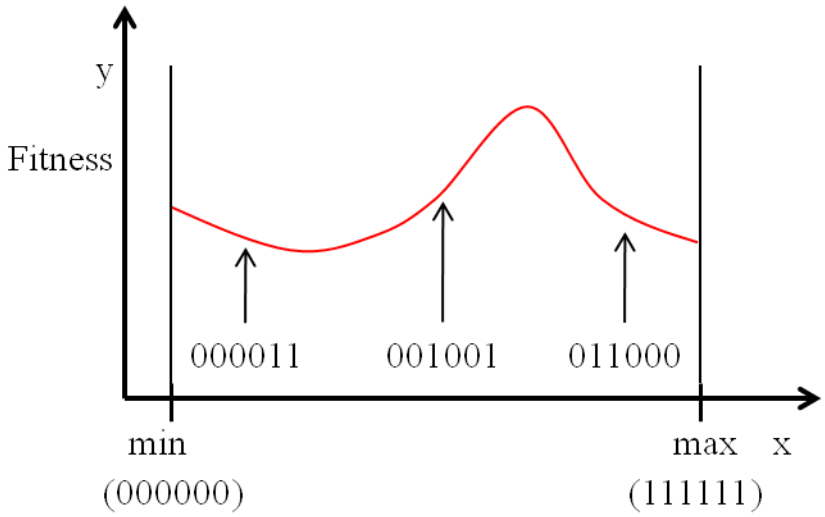


Abbildung 1: Initialisierung des Lösungsraumes; eindimensionales Beispiel
Der Eindimensionale Lösungsraum wird durch den minimalen bzw. maximalen Wert, den der Parameter annehmen kann, begrenzt. Dargestellt in rot ist die Fitness, von der das Optimum in dem Lösungsraum gesucht wird. Von jedem absoluten Wert in diesem Lösungsraum kann ein Bitstring errechnet werden, der hierbei aus sechs Bits besteht.

Zu jedem Punkt im Lösungsraum lässt sich die Fitness berechnen, wodurch alle Individuen miteinander verglichen werden können. Dies ermöglicht eine spätere Selektion der besten Individuen für die Nachfolgepopulation. Eine Nachfolgepopulation wird zunächst durch kopieren zufällig ausgewählter Individuen der Vorgängerpopulation[2] und anschließendem verändern der Individuen erzeugt. Die Individuen werden dabei paarweise ausgewählt, um eine Rekombination durchführen zu können. Sind zwei Individuen ausgewählt, wird zufällig entschieden ob Mutationen an den jeweiligen Bitstrings durchgeführt werden und ob ein Ein-Punkt Crossover zwischen beiden Bitstrings durchgeführt wird. Falls ein Crossover stattfindet wird stochastisch ein Crossoverpunkt bestimmt. Das Crossover spielt bei den Genetischen Algorithmen die größte Rolle, um eine optimale Lösung zu generieren. Bei einer Mutation findet eine Inkrementierung eines einzelnen Bits des Bitstrings statt, wobei sich der Wert des Bitstrings je nach Position des Bits um einen sehr kleinen oder auch sehr großen Betrag ändern kann. Die Mutation ermöglicht es ergänzend dazu den Lösungsraum sorgfältiger zu durchsuchen, da das Crossover keine Möglichkeit hat Teile einer Lösung gezielt zu verändern, sondern nur vorhandene Teile austauschen kann. Abschließend werden die Fitnesswerte der neu erstellten Individuen berechnet, selektiert und der Nachfolgepopulation zugeordnet. Bei der Selektion werden die Individuen mit der schlechtesten Fitness aussortiert, bis die Nachfolgepopulation die Größe der Vorgängerpopulation erreicht ist. [Goldberg, 1989; Holland, 1992]

[2] Auch Elternpopulation genannt.

1.2.3 Stärken und Schwächen des Algorithmus als Optimierungsverfahren

Durch eine Suche im gesamten Lösungsraum, funktioniert der Algorithmus gut bei der Suche nach dem globalen Optimum. Zusätzlich ist der Algorithmus relativ einfach, auch für komplexe Probleme, zu implementieren, wobei eine Anpassung der Parameteranzahl jederzeit ohne großen Aufwand möglich ist. [Nissen 1997]

Über generelle Verbesserungen der Parametereinstellungen des Algorithmus kann hingegen kaum eine Aussage getroffen werden, weil sich die unterschiedlichen Modifikationen durch den hohen stochastischen Anteil an dem Algorithmus schlecht miteinander vergleichen lassen. Zusätzlich sind die Parametereinstellungen des Algorithmus auch von der Struktur der Problemstellung abhängig. Bei dem Algorithmus kann keine Aussage über die Dauer der jeweiligen Suche getroffen werden.

1.3 Genutzte Datenbanken

1.3.1 MINT

Die *Molecular INTeraction* Datenbank (MINT) ist eine öffentliche Ansammlung an Protein-Protein Interaktionen (PPI), die in von Fachleuten geprüften Journalen aufgeführt sind. Die Datenbank enthält nach einem stetigen Wachstum über die Jahre im Oktober 2012 etwas mehr als 241.000 binäre Interaktionen an denen insgesamt 35.379 Proteine (von *Homo sapiens*: 8626) beteiligt sind. Die MINT Datenbank nutzt die Molekulare Interaktions Onthologie der *Proteomics Standard Initiative* (PSI-MI). PSI-MI Formate sind ein Standard zum Repräsentieren von molekularen Interaktionsdaten, die in über 30 Datenbanken implementiert, von Software Tools unterstützt und weitestgehend akzeptiert werden. Dieser PSI-MI Standard erlaubt eine bessere Kooperation zwischen öffentlichen Datenbanken aus der das *International Molekular Exchance* (IMEx) Konsortium [Orchard et al., 2012] hervorging. Damit ist es möglich mit Daten von verschiedenen Datenbanken zu arbeiten oder diese zu kombinieren, weil alle im gleichen Format abgelegt sind. [Licata et al., 2011]

Die MINT Datenbank war eine der ersten PPI Datenbanken, die jeder Interaktion einen Wert assoziierte, um die Zuverlässigkeit der Interaktion abschätzen zu können. Dafür wurde das Konzept des integrierten, unterstützenden Nachweises *y* definiert, dass die gewichtete Summe der *j* Manuskripte der gegebenen Interaktionen darstellt und dass eine Wichtung der „Anerkennung in der Community/Vertrauen der Community“ mit berücksichtigt.

$$y = \sum_j S_j R_j$$

Die Wichtung jedes unterstützenden Manuskriptes $S_i R_i$ wird durch multiplizieren zweier Koeffizienten erreicht, die beide im Intervall [0,1] variieren und die Aussagekraft des Experimentellen Nachweises (*S*) bzw. die Anerkennung und das Vertrauen der wissenschaftlichen Community (*R*) wiederspiegeln. Um *S* und *R* zu bestimmen, wurden zuerst *s* und *r* berechnet, die definiert sind als:

$$s = \sum_i e_i$$

r = normalisierte Zitierungen

Dabei ist *e* ist ein Koeffizient, der zwischen direkten Interaktionen (e = 1), experimentellen Ergebnissen ohne eindeutige Hinweise auf direkte Interaktionen, wie z.B. co-ip, Pulldown, etc. (e = 0,5) und Kolokalizationen (e = 0,1) unterscheidet. Im Gegenzug ist *r* das Verhältnis zwischen der Anzahl an Zitaten, erhalten durch das Manuskript gemäß Google Scholar (ergänzt um 20; begründet durch Vertrauen von der Community) und der Anzahl an unabhängigen im Manuskript beschriebenen Interaktionen. Ein angleichen von *s* und *r* an das Intervall [0,1] wird durch nachstehende Funktion umgesetzt, sodass *S* und *R* entstehen.

$$S = 1 - a^{-s}$$
$$R = 1 - b^{-r}$$

Dabei wurden *a* und *b* empirisch auf 1,2 gesetzt. Für einige Interaktionen ist kein Score angegeben. [Licata et al., 2011]

1.3.2 IntAct

Die IntAct ist eine open-source Datenbank, die Daten von molekularen Interaktionen enthält, die entweder aus der Literatur entnommen oder die direkt in die Datenbank hinzugefügt werden. Im Oktober 2012 enthält die Datenbank etwas mehr als 300.000 Interaktionen zwischen mehr als 63.000 Interaktoren[3], die aus über 5500 Publikationen entnommen wurden. Die IntAct Datenbank ist ebenfalls ein aktives Mitglied des IMEx Konsortiums. Somit sind die Daten unter Anderem in dem PSIMITAB Format abgelegt. Jeder Eintrag in der IntAct ist von Fachleuten geprüft (*senior Kurator*) und wird erst freigegeben, wenn dieser Kurator ihn akzeptiert. Abschließend, zum Release der Daten, wird der Autor von jeder Publikation kontaktiert und gebeten, auf die Darstellung

[3] Interaktoren sind mehrheitlich Proteine (definiert durch UniProtKB [Boutet et al., 2007]), aber auch kleine Moleküle (definiert durch ChEBI [Degtyarenko et al., 2008, de Matos et al, 2009]) und Gene (definiert durch Ensembl [Flicek et al., 2012]).

ihrer Daten Stellung zu nehmen. Sollte der Autor eventuelle Fehler feststellen, erfolgen abschließende manuelle Updates. [Kerrien et al., 2012]

Die Bewertung der Interaktionen erfolgt bei der IntAct durch den MIscore (*Molecular interactions confidence Score*).

$$S_{MI} \equiv MIscore \,||\, S \in [0,1]$$

Dieser ist im Intervall [0,1] normalisiert und berücksichtigt die Anzahl an Publikationen in denen über die Interaktion berichtet wurde, die Methode mit der die Interaktion ermittelt wurde und den Typ der Interaktion, der ermittelt wurde. Jeder von diesen drei Variablen wird durch jeweils einen Teil-Wert im Intervall [0,1] repräsentiert:

$$S_{[p,m,t]} \equiv Teil - Wert \,||\, S \in [0,1]$$
$$S_P \equiv Publikationenwert$$
$$S_M \equiv Methodenwert$$
$$S_T \equiv Typenwert$$

Zusätzlich wird der Einfluss jeder dieser drei Variablen durch einen Wichtungsfaktor bestimmt:

$$K_{[p,m,t]} \equiv Wichtungsfaktor \,||\, K \in [0,1]$$
$$K_p \equiv Wichtungsfaktor\ des\ Publikationenwertes;\ Voreinstellung: K_p = 1$$
$$K_m \equiv Wichtungsfaktor\ des\ Methodenwertes;\ Voreinstellung: K_m = 1$$
$$K_t \equiv Wichtungsfaktor\ des\ Typenwertes;\ Voreinstellung: K_t = 1$$

Nachfolgend werden die drei verschiedenen Teil-Werte betrachtet. Der Publikationenwert berücksichtigt die verschiedenen Publikationen in denen die Interaktion ermittelt wurde:

$$S_p \equiv Publikationenwert \,||\, S \in [0,1]$$
$$S_p = log_{b+1}(n+1)$$
$$n \equiv Anzahl\ Publikationen, die\ über\ die\ Interaktionen\ berichten \,||\, n \in \mathbb{N}$$
$$b \equiv Anzahl\ an\ Publikationen\ für\ maximalen\ Wert;\ Voreinstellung\ b = 7$$

Durch die Voreinstellung der maximal benötigten Publikationen auf sieben bekommen alle Einträge mit sieben oder mehr Publikationen den Wert 1. Der Methodenwert berücksichtigt die Vielfalt an Annotationen, die für eine Interaktion bekannt sind:

$$S_m \equiv Methodenwert \,||\, S \in [0,1]$$
$$S_m(cv_i) = log_{b+1}(a+1)$$
$$a = \sum(scv_i * n_i)$$
$$b = a + \sum(Max(Gscv_i))$$

Dabei ist *cv* die Methode mit der die Interaktion ermittelt wurde. Validierte Methoden sind in der MI Ontologie [Cote et al., 2006] als Kinderknoten des Ontologie Terms „*interaction detection method*“ definiert. Der Wert, der zu einer Methode definiert wurde, ist mit *scv* beschrieben, wobei *scv* ∈ [0,1]. Die voreingestellten Werte für *scv* sind in nachstehender Tabelle 1 aufgeführt. Die Anzahl, wie oft ein Ontologie Term aufgeführt wird, ist mit *n* gekennzeichnet und *Gscv* beschreibt eine Gruppe an Werten.

Tabelle 1: Voreingestellte *scv*-Werte des Methodenwertes
An dieser Stelle sind die voreingestellten Werte (*scv*) der experimentellen Methoden (*cv*) aufgeführt. Dabei diente die Kinderknoten des Ontologie Terms „interaction detection method“ der MI Ontologie als Grundlage. Existiert kein Wert für einen Ontologie Term, der eine Methode beschreibt, wird der Wert des nächstgelegenen Elternknoten verwendet. Existiert kein derartiger Elternknoten wird der Wert 0,05 gewählt. [Smith et al., 2007]

Bezeichnung	Wert	Ontologie ID der Methode	Methodenbeschreibung
scv_1	1,00	MI:0013	Biophysisch
scv_2	0,66	MI:0090	Protein Komplementations Assay
scv_3	0,10	MI:0254	genetische Interferenz
scv_4	0,10	MI:0255	posttranskriptionale Interferenz
scv_5	1,00	MI:0401	biochemisch
scv_6	0,33	MI:0428	Assoziationstechnik
scv_7	0,05	unbekannt	unbekannt

Der Typenwert berücksichtigt ebenfalls die Vielfalt an Annotationen, die für eine Interaktion bekannt sind:

$$S_t \equiv Typenwert \,||\, S \in [0,1]$$
$$S_t(cv_i) = log_{b+1}(a+1)$$
$$a = \sum(scv_i * n_i)$$
$$b = a + \sum(Max(Gscv_i))$$

Hierbei ist *cv* der Typ der entsprechenden Interaktion. Validierte Interaktionstypen sind in der MI Ontologie [Cote et al., 2006] als Kinderknoten des Ontologie Terms „*interaction tpye*“ definiert. Der Wert, der zu einem Interaktionstyp definiert wurde, ist mit *scv* gekennzeichnet, wobei

scv ∈ [0,1] ist. Die voreingestellten Werte für *scv* sind in nachstehender Tabelle 2 aufgeführt. Die Anzahl, wie oft ein Ontologie Term aufgeführt wird, ist mit *n* gekennzeichnet und *Gscv* beschreibt eine Gruppe an Werten.

Tabelle 2: Voreingestellte *csv* Werte des Typenwertes
An dieser Stelle sind die voreingestellten Werte (*scv*) der Interaktionstypen (*cv*) aufgeführt. Dabei dienten die Kinderknoten des Ontologie Terms „interaction type" der MI Ontologie als Grundlage. Existiert kein Wert für einen Ontologie Term, der einen Interaktionstyp beschreibt, wird der Wert des nächstgelegenen Elternknoten verwendet. Existiert kein derartiger Elternknoten wird der Wert 0,05 gewählt. [Smith et al., 2007]

Bezeichnung	Wert	Ontologie ID des Interaktionstyps	Typenbeschreibung
scv_1	0,10	MI:0208	Genetische Interaktion
scv_2	0,33	MI:0403	Kolokalisation
scv_3	0,33	MI:0914	Assoziation
scv_4	0,66	MI:0915	Physische Assoziation
scv_5	1,00	MI:0407	Direkte Interaktion
scv_6	0,05	unbekannt	unbekannt

Zusammen ergibt sich nachstehende Formel für den MIscore:

$$S_{MI} = \frac{K_p * S_p(n) + K_m * S_m(cv) + K_t * S_t(cv)}{K_p + K_m + K_t}$$

[Braun et al., 2009]

1.3.3 SCOPPI

Die SCOPPI (structural classification of protein-protein interfaces) ist eine umfassende Datenbank, die Domänen-Interaktionen klassifiziert und kommentiert, die aus allen bekannten Proteinstrukturen abgeleitet sind. Die SCOPPI verwendet die SCOP Domänen Definitionen [Andreeva et al., 2007] und ein Distanzkriterium[4], um Domänenschnittstellen zu bestimmen. Durch eine Methode, die auf Multisequenzalignments und strukturellen Alignments der SCOP Familien beruhen, ist es der SCOPPI Datenbank möglich, eine umfassende, geometrische Klassifizierung von Domänenschnittstellen darzubieten. Verschiedene Eigenschaften der Schnittstellen, wie Anzahl, Art und Position der interagierenden Aminosäuren, Konservierung, Schnittstellengröße und dauerhafte oder vorübergehende Art der Wechselwirkung werden zusätzlich betrachtet. Die Proteine in der SCOPPI sind mit der Gene Ontologie annotiert und die Ontologie kann dazu genutzt werden in der SCOPPI zu navigieren. Die Kombination von Multisequenzalignments und strukturellen Alignments von den SCOP Familien erlaubt eine gründliche Klassifizierung von Bindestellen zu anderen Domänen.

[4] Das Distanzkriterium basiert auf der PSIMAP [Dafas et al., 2004; Kim et al., 2004] Methode.

Für alle alignierten Domänen einer SCOP Familie werden die Bindestellen zu anderen Domänen *faces* genannt, die entsprechend den strukturell und sequenziell überlappenden Merkmalen in *face types* klassifiziert wurden. Eine SCOP Familie kann viele *face types* haben, abhängig von ihrer Vielfalt an Bindestellen. Zwei interagierende *face types* stellen einen *interface type* dar. Angelehnt an andere Definitionen von Schnittstellen [Tsai et al., 1996] definiert die SCOPPI zwei Domänen als interagierend, wenn sie wenigstens fünf Residue-Residue Kontakte innerhalb von 5 Å aufweisen [Gong et al., 2005]. [Winter et al., 2006]

1.3.4 GeneCards

Die GeneCards ist ein umfassendes, autoritäres Kompendium über annotierte Informationen von menschlichen Genen, dessen Inhalt durch automatisches Datamining aus über 80 digitalen Quellen erfasst wird. Das Resultat ist eine webbasierte Karte für jeden der über 73.000 Einträge an menschlichen Genen, die folgenden Kategorien umfassen: Proteincodiert, Pseudogen, RNA-Gene, genetische Loci, Cluster und nicht kategorisierte. Ein Schwerpunkt liegt darin Gen-Sets zu analysieren, die GeneCards einzigartige Fülle an kombinatorischen Annotationen nutzen. Ein für diese Arbeit wichtiger Teil der GeneCards ist die Auflistung von interagierenden Proteinen. Diese Liste ist im „Pathways & Interactions“ Bereich ersichtlich, der in der Detailansicht für ein Protein betrachtet werden kann. Dabei ist in jeder Zeile der Tabelle ein interagierendes Protein aufgelistet, dass entsprechend der Informationen an Protein-Protein Interaktionen aus der UniProtKB [UniProt Consortium, 2012], der EBI-IntAct, der STRING [von Mering et al., 2003], MINT und I2D [Brown & Jurisica, 2005; Brown & Jurisica, 2007] zusammen getragen und tabellarisch dargestellt wird. Diese Tabelle beinhaltet in der ersten Spalte (GeneCard) Links zu den GeneCards Einträgen der einzelnen Interaktionspartner sowie in der zweiten Spalte (External ID(s)) Links zu den Einträgen der Interaktionspartner auf der UniProtKB und/oder Ensembl. In dieser Spalte ist mittels Fußnoten zusätzlich angegeben, von welcher Datenbank der jeweilige Interaktionspartner vermittelt wurde. Diese Fußnoten sind in nachstehender Tabelle 3 aufgeschlüsselt. Die dritte Spalte (interaction Details) beinhaltet Links zu den Interaktionsdatenbanken, von welchen die Daten abgerufen wurden. [Safran et al., 2010] Wird im Folgenden davon gesprochen, dass ein Protein in der GeneCards Datenbank annotiert ist, so ist damit das auftauchen des Proteins in der Tabelle der Interaktionspartner von dem aktuell zu untersuchenden Protein gemeint.

Tabelle 3: Fußnotenbeschreibung der Interaktionspartner auf der GeneCards
Hierbei ist beschrieben, auf welche Quellen die Interaktionspartnerliste der Genecards zurückzuführen ist.

Fußnote	Beschreibung
1	Der Kommentarbereich auf der UniProtKB
2	Die Seite aller Wechselwirkungen zwischen den beiden Proteinen oder allen Experimenten die sie unterstützen in der MINT-Datenbank
3	Die Seite aller Interaktionen mit dem Interaktionspartner der I2D Datenbank
4	Das Interaktion Netzwerk der Interaktionspartner der String Datenbank

Das GeneCards Kompendium beinhaltet weitere interessante Features, die an dieser Stelle nicht betrachtet wurden, weil sie in dieser Arbeit keine Rolle spielen.

1.3.5 PIPs – Human Protein-Protein Interaction Prediction

Die PIPs Datenbank ist eine Ressource, die sich mit dem Studium von Protein-Protein Interaktionen von menschlichen Proteinen beschäftigt. Die Vorhersage der Protein-Protein Interaktionen zwischen menschlichen Proteinen wurde durch ein bayessches Framework, unter Berücksichtigung einer Kombination an individuellen Proteineigenschaften, von denen bekannt ist, dass sie auf Interaktion hinweisen, untersucht. Die sieben individuellen Proteineigenschaften sind in Tabelle 4 aufgeführt. Die „Module“ Expression, Orthologie, Kombiniert und Erkrankung können Wahrscheinlichkeiten (LR, engl. *likelihood ratio*) von Interaktionen unabhängig voneinander berechnen und werden somit als Gruppe A bezeichnet. Das Produkt dieser vier LRs der Gruppe A Module wird als *Preliminary Score* bezeichnet. Das Modul Transitiv berücksichtigt die lokale Topologie des Netzwerkes, dass durch die Module der Gruppe A vorhergesagt wurde und benötigt daher für die Berechnung des eigenen LRs ($LR_{Transitiv}$) die Fertigstellung aller Module der Gruppe A. Die finale Wahrscheinlichkeit der Interaktion wird aus dem *Preliminary Score* und dem $LR_{Transitiv}$ errechnet. Kann keine lokale Topologie erstellt werden, so wird der *Preliminary Score* als finale Wahrscheinlichkeit verwendet. [Scott & Barton, 2007; McDowall et al., 2009]

Im Folgenden wird auf die einzelnen Module näher eingegangen. Für das Expressionsmodul wurde der GDS596 Expressionsdatensatz der Gene Expression Omnibus [Barrett et al., 2005] genutzt, der Genexpressionsprofile von 79 physiologisch normalen Geweben aus verschiedenen Quellen untersucht [Su et al., 2004]. Von allen möglichen Paaren, der eindeutigen Transkripte, wurden die Pearson Korrelationskoeffizienten bestimmt, womit die Koexpression beschrieben wird. [Scott & Barton, 2007]

Tabelle 4: Bestandteile des PIPs bayesschen Frameworks

Modul	Proteineigenschaft	Datenquelle
Expression	Expression	GDS596 des Gene Expression Omnibus
Orthologie	Orthologie	Datenbanken: InParanoid, BIND, DIP, GRID
Kombiniert	Lokalisation	PSLT Vorhersage
	Domänen Kookkurrenz	InterPro, Pfam
	PTM Kookkurrenz	HPRD, UniProt
Erkrankung	Erkrankung	VLS2 Vorhersage
Transitiv	Transitiv	

Für das Orthologiemodul wurden Orthologiekarten von der InParanoid Datenbank [O'Brien et al., 2005] zwischen Mensch und Hefe, Mensch und Fliege bzw. Mensch und Wurm bezogen. Zusätzlich wurde ein Interaktionsdatensatz aus den Datenbank BIND [Alfarano et al., 2005], DIP [Salwinski et al., 2004] und GRID [Breitkreutz et al., 2003] erstellt. Die Orthologie Interaktionsdaten wurden für jedes Protein-Interaktionspaar des Menschen, dass ein Interaktionspaar in einem der drei anderen Organismen hat, gemäß den Werten aus der InParanoid Datenbank, klassifiziert. [Scott & Barton, 2007]

Das Kombinierte Modul enthält drei einzelne Proteineigenschaften. Für die Subzellulare Lokalisation wird das PSLT (*Protein Subcellular Localization Tool*) [Scott et al., 2004] genutzt, dass die Proteine eines Proteinpaare in vier Gruppen einteilt: gleiches Zellkompartiment, benachbartes Zellkompartiment, nicht benachbartes Zellkompartiment und keine Vorhersage möglich [Scott & Barton, 2007]. Für die Domänen Kookkurrenz wurde der *chi-square* Test als eine Messung der Wahrscheinlichkeit der Kookkurrenz von spezifischen InterPro Domänen und Motiven von Proteinpaaren genutzt [Mulder et al., 2005]. Zusätzlich wurden Pfam [Finn et al., 2006] Domänenpaare genutzt, von denen von dreidimensionalen Strukturen bekannt ist, dass sie miteinander interagieren [Jefferson et al., 2007]. Verwendet wurde der daraus resultierende *chi-square* Wert. Die posttranslationalen Modifizierungen (PTM) von menschlichen Proteinen wurden von der UniProt [Wu et al., 2006] und der HPRD [Mishra et al., 2006] genutzt. Der Wert für die PTM wurde berechnet, indem die Wahrscheinlichkeit des gleichzeitigen Auftretens zweier PTMs in allen interagierenden Proteinpaaren geteilt durch die Wahrscheinlichkeit des separaten Auftretens von beiden PTMs. [Scott & Barton, 2007]

Für das Krankheitsmodul wurde ein Wert genutzt, der aus dem VSL2B Tool für jedes der interagierenden Proteine berechnet wurde. Für den finalen Wert wurden beide Einzelwerte addiert [Peng et al., 2006]. [Scott & Barton, 2007]

Das Transitive Modul arbeitet unter der Voraussetzung, dass zwei Proteine mit einer höheren Wahrscheinlichkeit miteinander interagieren, wenn sie gemeinsame Interaktionspartner haben. Dies wird berücksichtigt, indem um die potentielle Proteininteraktion ein lokales Netzwerk anhand der

Werte aus den Gruppe A Modulen erstellt und die Topologie dieses lokalen Netzwerkes rund um das Proteinpaar bewertet wird. Dafür wird für das Transient Modul ein eigener LR-Wert ($LR_{Transient}$) berechnet, der aus den *Preliminary Scores* der Nachbarn im Netzwerk, nicht aber aus dem *Preliminary Score* des betrachteten Proteinpaares selbst errechnet wird. Der $LR_{Transient}$ wird anschließend mit dem *Preliminary Score* des Proteinpaares verrechnet. [Scott & Barton, 2007] Wird im Folgenden davon gesprochen, dass ein Protein in der PIPs Datenbank annotiert ist, so ist damit eine vorhergesagte Interaktion des Proteins mit dem relevanten, zu untersuchendem Protein gemeint.

1.4 Untersuchte Proteine

Untersucht wurden folgende Proteine: STAT3 (*Signal transducer and activator of transcription 3*), STAT1 (*Signal transducer and activator of transcription 1*), CDK9 (*Cyclin-dependent kinase 9*) und BMI1 (*Polycomb complex protein BMI-1*).

1.4.1 STAT-Familie

Im Folgenden wird speziell auf STAT1 und STAT3 eingegangen, da beide ein wichtiger Bestandteil dieser Arbeit sind.

STAT-Proteine sind Transkriptionsfaktoren, die über den JAK-STAT-Signalweg am Immunsystem, am Zellwachstum und der Proliferation beteiligt sind. Die sieben Mitglieder der Säugetier STAT-Familie (STATs 1, 2, 3, 4, 5a, 5b, und 6) umfassen einen Bereich von 750 bis 900 Aminosäuren und weisen verschiedene konservierte Domänen, insbesondere der SH2 Domäne, auf. In ruhenden Zellen verweilen die STAT-Proteine weitestgehend im Zytoplasma als inaktive Homodimere [Mertens et al., 2006]. [Schindler et al., 2007]

Das STAT3-Protein vermittelt zelluläre Reaktionen auf Wachstumsfaktoren und Interleukine, wie dem Interleukin (IL)-6-Typ Zytokin. Die Signale dieses IL-6 werden durch das Glykoprotein gp130 in die Zielzellen übermittelt, wobei das gp130 durch die JAK (Januskinasen) aktiviert wird [Veverka et al., 2012]. Die Aktivierung bewirkt eine Phosphorylierung von Tyrosin 705 an den STAT3-Monomeren, die daraufhin antiparallele Homodimere oder Heterodimere mit STAT1 bilden, indem die SH2 Domänen eines STAT-Proteins an das Phosphotyrosin des anderen STAT-Proteins bindet. Dies ermöglicht eine Verlagerung aus dem Zytoplasma in den Zellkern, in dem das dimerisierte Molekül als Transkriptionsfaktor mit anderen Transkriptionsfaktoren interagiert und somit unter Anderem eine wichtige Rolle bei dem Zellwachstum, der Zelldifferenzierung und der Apoptose spielt [Wegenka et al., 1994]. Während anschließend STAT1 durch eine

Tyrosin Dephosphorylierung inaktiviert und zurück ins Zytoplasma gelangt, [Haspel & Darnell, 1999] haben sich verschiedene Studien in den letzten Jahren mit den posttranslationalen Interaktionspartnern von STAT3 beschäftigt. Beispielsweise phosphorylieren Serin Kinasen das Serin 727 dieses Proteins; einige Lysine von STAT3 werden von Acetyltransferasen acetyliert oder es binden Coaktivatoren, die das Transkriptions-Aktivierungspotential von STAT3 erhöhen oder verringern [Rodel et al., 2000; Yang et al, 2005; Dabir et al., 2009]. Weiterhin kann STAT3 selbst als Coaktivatorprotein agieren, indem es mit anderen Transkriptionsfaktoren interagiert [Proietti et al., 2011]. Im Jahr 2009 wurde STAT3, das ursprünglich als Transkriptionsfaktor im Zellkern eingeordnet wurde, in anderen Zellkompartimenten nachgewiesen. Mutanten von STAT3, deren Tyrosin nicht phosphoryliert werden kann oder das nicht an die DNA binden kann, unterstützt in Mitochondrien die Transformation des Proto-Onkogens Ras. [Gough et al., 2009] Zusätzlich interagiert es mit Komponenten des Zytoskeletts, wie Stathmin. Dabei reguliert es die Mikrotubuli Dynamiken bei dem Transport von T-Zellen. [Verma et al., 2009] Diese Beobachtungen zeigen das STAT3, neben der hohen Bedeutung als Transkriptionsfaktor, auch eine wichtige Rolle für die Mechanismen der Signalübertragung spielt. Somit ist STAT3 mit einer Vielzahl an Proteinen verknüpft, die in unterschiedlichsten Prozessen beteiligt sind. Auch seine Beteiligung an verschiedensten Krankheitsbildern, wie Prostatakrebs, Leukämie, Malignes Lymphom (Lymphdrüsenkrebs), multiples Myelom [Stepkowski et al., 2008] und dem Hyper IgE Syndrom [Holland et al., 2007; Minegishi et al., 2007], unterstreicht die Wichtigkeit eines umfassenden Bildes des Interaktomes von STAT3. [Schindler et al., 2007]

Das STAT1-Protein spielt eine zentrale Rolle bei der Vermittlung von physiologischen und therapeutischen Wirkungen von Interferonen auf menschliche Zellen. Daher weisen Menschen mit Mutationen in STAT1 eine Anfälligkeit für bakterielle und virale Infektionen auf. Außerdem kann eine Mutation mit Funktionszugewinn in STAT1 für chronisches *mucocutaneous candidiasis*[5] (CMC) verantwortlich sein [Takezaki et al., 2012]. Interessanter weise scheinen die Zielgene von STAT1 Entzündungen zu fördern und Proliferation entgegen zu wirken. Dies steht im Kontrast zu den entzündungshemmenden und profilerationsfördernden Aktivitäten von STAT3. Durch die Fähigkeit mehrerer Zytokine sowohl STAT1, als auch STAT3 zu aktivieren, ist dieses System in einem Ausgewogenen zustand. [Schindler et al., 2007]

1.4.2 CDK9

Cyclin-dependent kinase 9 (CDK9) ist eine Proteinkinase, die in der Regulation der Transkription involviert ist. Sie ist mit ihrer Cyclin T Untereinheit Teil des P-TEFb (*positive transcription*

[5] CMC ist eine Gruppe an hauptsächlich Immunschwächekrankheiten, die durch chronische oder wiederkehrende Candida Infektionen der Haut, Nägel und des Mundrachenraums charakterisiert wird.

elongation factor b) Komplexes, der die Elongation der Transkription stimuliert, aber auch Funktionen bei der mRNA Verarbeitung, dem mRNA Export und der Modifikation von Kotranskroptionellen Histonen übernimmt. CDK9 bindet auch an Cyclin K, wobei die Funktion dieses CDK9-cyclin K Komplexes nicht ganz geklärt ist. CDK9 ist direkt in Aufgaben im Bereich der Aufrechterhaltung der Unversehrtheit des Genoms involviert, wobei diese Aktivität auf den CDK9-cyclin K Komplex beschränkt ist. Der Abbau von CDK9 oder dessen Cyclin K, aber nicht Cyclin T, Untereinheit beeinträchtigt die sich replizierenden Zellen während des Zellzyklus in Bezug auf Stress bei der Replikation, das zu spontanen DNA Schäden führen kann. CDK9-cyclin K interagiert auch mit Proteinen, die Schäden an der DNA feststellen oder Schäden an der DNA reparieren. Als Reaktion auf Stress, der bei der Replikation entsteht, lagert sich CDK9 an Chromatin an und limitiert den Betrag an einsträngiger DNA. Die direkte Rolle von CDK9-cyclin K in Stoffwechselwegen, die Erhaltung der Unversehrtheit des Genoms als Reaktion auf Stress bei der Replikation, scheint evolutionär konserviert zu sein. Die präzise Replikation des Genoms und die kontinuierliche Überwachung der Unversehrtheit ist essentiell für das Überleben der Zellen und dem Vermeiden von verschiedenen Krankheiten, einschließlich Krebs. [Yu & Cortez, 2011] Dieses Protein wurde in dieser Arbeit ausgewählt, weil es in den Zytokin induzierten Transkriptionsnetzwerken, wie beispielsweise dem IL-6 induziertem STAT3 Signalweg, involviert ist. [Hou et al., 2007]

1.4.3 BMI1

Das *Polycomb complex protein BMI-1* ist Bestandteil von der Polycomb Gruppe (PcG), dessen Proteine eine Hemmung der Transkription von hunderten von Genen aufrechterhalten, die in der Entwicklung, Signalübertragung oder in Krebs involviert sind. Biochemische Studien in *Drosophila* haben gezeigt, dass PcG Proteine in zwei Klassen an Proteinkomplexen eingeteilt werden können: dem PCR1 und dem PCR2 (*Polycomb repressive complexes 1*). Dabei besteht das PCR1 aus vier Untereinheiten: Pc (*Polycomb*), Sce (*Sex combs extra*), Ph (*Polyhomeotic*) und Psc (*Posterior sex combs*). BMI1 stellt dabei unter Anderem ein ontologes Protein aus Säugetierzellen zu dem Psc dar. Der PCR1 Komplex modifiziert posttranslational Residuen an Histonen und Protein Kinase 2 bindet an PCR1, wodurch der Polycomb Komplex zusätzlich eine Kinase Aktivität bekommt. [Vandamme et al., 2011] BMI1 stellt bei der STAT3/STAT1 Interaktomanalyse eine negative Vergleichsprobe dar, da es funktionell mit diesem Komplex in Verbindung steht, aber nicht Aktivierungsabhängig ist.

2 Zielstellung

Im Rahmen dieser Masterarbeit wurde das Interaktom der Proteine STAT3, STAT1, BMI1 und CDK9 in humanen embryonalen Nierenzellen (HEK 293T, engl. *Human Embryonic Kidney 293T cells*) mit einer auf Affinitätsmassenspektrometrie basierenden Strategie untersucht, in der stabile Isotopenmarkierung durch Aminosäuren in Zellkultur (SILAC), *in situ* Biotinylierung der vier ausgewählten Proteine, Affinitätsanreicherung und massenspektrometrische Analyse (AP-MS) verbunden wurden. Den Schwerpunkt der Arbeit stellte die Optimierung der Datenauswertung dar. Zu diesem Zweck wurde eine Software entwickelt, die ein Protein-Protein Interaktionsnetzwerk aus Interaktionsdatenbanken um das jeweils zu untersuchende Protein erstellt und mit Hilfe von einer Meta-Datenbank und dem Protein-Protein Interaktionsnetzwerk die signifikanten Bindungspartner der Analyse selbstständig ermittelt.

Für ein umfassendes Bild des STAT3/STAT1 Interaktoms sollten zur Detektion der Interaktionspartner auch CDK9 und BMI1 untersucht werden. BMI1 stellt dabei die negative Vergleichsprobe dar, da es funktionell mit dem STAT3/STAT1 Komplex in Verbindung steht, aber weder mit STAT1, noch mit STAT3 direkt interagiert sowie nicht aktivierungsabhängig reagieren sollte. CDK9 ist hingegen ein bekannter STAT-Bindungspartner [Hou et al., 2007]. Es sollte z.T. mit STAT1 oder STAT3 gemeinsame Bindungspartner aufweisen. Ziel dieser Arbeit war zunächst das Erstellen einer automatischen Auswertung der durch AP-MS gewonnenen Daten zur Generierung verlässlicher Listen potentieller Bindungspartner. In einem zweiten Schritt sollten die ermittelten Bindungspartner mit Dreieck-Netzwerk-Motiven und komplementären Daten nachprozessiert bzw. nachevaluiert werden. Die Performance des Ansatzes wurde durch den Vergleich mit der PIPs Datenbank verglichen und evaluiert.

3 Materialien und Methoden

3.1 Genutzte Software

Programmbezeichnung	Version	Bemerkung
Proteom Discoverer	1.2	Thermo Fisher Scientific GmbH, Dreieich, Deutschland
Eclipse *Juno*	4.2.0	Entwicklungsumgebung für Java
WindowBuilder	3.7	Plugin für Eclipse für GUI-Konstruktion
AmiGO Webservice	1.7	Release 02.12.2009
BLAST	2.2.26	standalone des NCBI FTP Servers (Datei: „ncbi-blast-2.2.26+-x64-win64.tar.gz“)

3.2 Anfertigung und Messung der Proben

Alle Experimente wurden in biologischen Triplikaten durchgeführt. Die Zellkultur der humanen embryonalen Nierenzellen (HEK 293T), die stabile Isotopenmarkierung durch Aminosäuren in Zellkultur (SILAC), die anschließende *in situ* Biotinylierung der Proteine STAT3, STAT1, BMI1 und CDK9 und die nachfolgenden Affinitätsaufreinigungen wurden von Gabriele Pfeiffer und Conny Blumert (AG Prof. Friedemann Horn, Universität Leipzig) durchgeführt. Dabei wurden die Zellenkulturen des schweren Ansatzes zur Detektion aktivierungsabhängiger Bindungspartner mit Wachstumsfaktoren stimuliert. Weitere Details der drei SILAC Ansätze sind in nachstehender Tabelle 5 dargestellt. Anschließend wurden ein Streptavidin Pulldown und die Aufreinigung dieser Proben durchgeführt. Die drei Ansätze wurden gemischt und die Proteine wurden gelelektrophoretisch mittels 1D-SDS-PAGE getrennt. Im Anschluss erfolgte durch Jacqueline Kobelt (AG Kalkhof, Proteomics Department des UFZ) ein in-Gel-Verdau mittels Trypsin. Die Proben aller drei biologischen Replikate der Pulldowns von CDK9, BMI1, STAT3 und STAT1wurden mittels nano-HPLC/nano-ESI-MS/MS sowohl an einem LTQ Orbitrap XL ETD sowie einem LTQ Orbitrap Velos Massenspektrometer durch Dr. Stefan Kalkhof (Proteomics Department des Helmholtz-Zentrum für Umweltforschung – UFZ, Leipzig) gemessen.

Tabelle 5: Übersicht SILAC-Versuche
Diese Tabelle veranschaulicht die verschiedenen Ansätze mit den verwendeten Aminosäuren und den jeweiligen Wachstumsfaktoren, mit denen die einzelnen Zellen stimuliert wurden. Alles wurde in HEK-293T Zellen durchgeführt und von jedem Ansatz wurden je Protein drei Replikate angefertigt.

Bezeichnung	Leichter Ansatz	Mittlerer Ansatz	Schwerer Ansatz
STAT3	GFP-Bio, unstimuliert	STAT3-Bio, unstimuliert	STAT3-Bio, EPO
STAT1	GFP-Bio, unstimuliert	STAT1-Bio, unstimuliert	STAT1-Bio, INFγ
BMI1	GFP-Bio, unstimuliert	BMI1-Bio, unstimuliert	BMI1-Bio, EPO
CDK9	GFP-Bio, unstimuliert	CDK9-Bio, unstimuliert	CDK9-Bio, EPO
verwendete Aminosäuren	$^{1}H_4\,^{12}C_6\,^{14}N_2$ L-Lysin $^{12}C_6{}^{14}N_4$ L-Arginin	$^{2}H_4\,^{12}C_6\,^{14}N_2$ L-Lysin $^{13}C_6{}^{14}N_4$ L-Arginin	$^{1}H_4\,^{13}C_6\,^{15}N_2$ L-Lysin $^{13}C_6{}^{15}N_4$ L-Arginin

3.2.1 Vorversuch

Die Bestimmung der Fitnessfunktion basierte auf einem Versuch, der ebenfalls von der AG Dr. Stefan Kalkhof, Proteomics Department des UFZ, Leipzig, in Zusammenarbeit mit der AG Prof. Friedemann Horn, Universität Leipzig, durchgeführt wurde. Dabei wurden ebenfalls die Zellkultur der humanen embryonalen Nierenzellen (HEK 293T), die stabile Isotopenmarkierung durch Aminosäuren in Zellkultur (SILAC), die anschließende *in situ* Biotinylierung des Proteins STAT3 und die nachfolgenden Affinitätsaufreinigungen von Gabriele Pfeiffer und Conny Blumert durchgeführt. Die Zellkulturen des schweren Ansatzes wurden dabei biotinyliert und zusätzlich mit EPO stimuliert. Der leichte Ansatz blieb unverändert. Durch eine voraus gegangene Zellfraktionierung konnten Messungen der Peptide, zuzüglich zu den Messungen des Gesamtzelllysat, auch im Zellkern und Zytosol durchgeführt werden. Diese Experimente wurden in biologischen Duplikaten durchgeführt. Ferner wurden als Kontrollexperiment vier Replikate mit einer Anreicherung von GFP [Tsien, 1998] erstellt, wobei das Gesamtzelllysat gemessen wurde.

3.3 Auswertung der AP-MS/MS Daten

Die gemessenen RAW-Daten wurden mittels Proteom Discoverer einer ersten Auswertung unterzogen. Dafür wurde ein Workflow mit den Komponenten *Spectrum Selector*, *Maskot*, *Event Detector* und *Precursor Ion Quantifier* angefertigt, deren jeweilige Parameter nachfolgend benannt werden, sofern diese von den Standardeinstellungen abweichen. Im *Spectrum Selector* wurde der Massenbereich des Spektrums auf 350 Da – 5000 Da und der minimal zählbare Peak auf eins festgelegt. Im *Maskot* Modul wurde die Taxonomie auf Trypsin und menschlichen Proteinen sowie die strenge FDR (engl. *false discovery rate*) auf 0,01, die lockere FDR auf 0,05, die Vorläufermassen-

toleranz auf 5 ppm und die Massentoleranz der Fragmente auf 0,5 Da gestellt. Es wurde keine durchschnittliche Vorläufermasse genutzt. Als dynamische Modifikation wurde *Acetyl (Protein N-term)* und als statische Modifikation wurde *Carboamidomethyl (C)* eingestellt. Die dynamischen Markierungen waren *13C(6) (R), 13C(6) 15N(4) (R), 13C(6) 15N(2) (R)* sowie *2H(4) (R)*. Für das Modul *Event Detector* wurden 4 ppm bestimmt.

Durch den Trypsin-Verdau werden bei der Messung keine vollständigen Proteine, sondern nur Peptide bestimmt, die mit dem Proteom Discoverer mittels SwissProt Datenbank den entsprechenden Proteinen zugeordnet wurden. Dabei verwendet der Proteom Discoverer die UniProtKB-AN). Durch die markierten Aminosäuren Lysin und Arginin kann bestimmt werden, aus welchem der drei Ansätze (leicht, engl. *light*, L; mittel, engl. *middle*, M; schwer, engl. *heavy*, H) die Proteine hervorgegangen sind. Der Proteom Discoverer bestimmt aus diesen Angaben die Verhältnisse der drei Ansätze (H/L, M/L, H/M) für jedes gemessene Protein, sofern das entsprechende Protein in den entsprechenden Ansätzen gemessen wurde. Diese Daten werden vom Proteom Discoverer für jedes Replikat erstellt und können in eine Textdatei[6] exportiert werden.

Zur abschließenden Auswertung wurde ein Parser konstruiert, der die Daten dieser Textdatei aufbereiten kann, wobei der Parser die Daten von jedem der drei Ansätze (H/L, M/L, H/M) getrennt betrachtet. Aus den in den Ansätzen enthaltenen Daten bestimmt der Parser zuerst die Anzahl an verwendbaren Replikaten von jedem Protein, weil es vorkommen kann, dass ein Protein bei der Messung in einigen Replikaten nicht enthalten ist. Als nächsten Schritt normiert der Parser die Verhältnisse jedes Proteins. Dabei wird der Mittelwert der Mediane der Verhältnisse berechnet, wodurch sich Mischfehler bei der Probenanfertigung und ein nicht 100 %-iger Austausch der markierten mit den natürlichen Aminosäuren korrigieren lassen. Anschließend wurden diese normierten Verhältnisse zur Basis 2 logarithmiert. Aus diesen normierten und logarithmierten Verhältnissen, sofern mindestens zwei verwendbare Replikate von dem betreffenden Protein vorhanden waren, wurden der Mittelwert, der T-Test und die Standardabweichung berechnet. Dabei wurde für die Ansätze H/L und M/L ein einseitiger, gepaarter T-Test gegen null und für den Ansatz H/M ein zweiseitiger, gepaarter T-Test gegen null berechnet. Dies hat den Hintergrund, dass bei dem Quotient zwischen Proteinen in stimulierten und Proteinen in unstimulierten Zellen sowohl eine Anreicherung als auch eine Abreicherung als wichtige Information gewertet werden kann. Ein weiterer Punkt sind einige Filteroptionen, die von dem Parser umgesetzt werden. Der erste Filter bezieht sich auf Proteine, die eine Domäne haben, die an Biotin binden kann. Dabei wurden fünf Proteine[7] mit Hilfe der UniProtKB Datenbank [Boutet et al., 2007] mit folgender Suchanfrage bestimmt: „"biotinyl binding" AND domain AND organism:human“. Eine weitere Filterfunktion

[6] Dabei ist darauf zu achten, dass nur die erste Ebene der Tabelle aus dem Proteom Discoverer exportiert werden muss.

[7] Diese hatten folgende UniProtKB-ANs: P11498, P05165, Q13085, O00763 und Q96RQ3.

bezieht sich auf Chaperone und Proteolysefaktoren. Diese Liste wurde anhand der Einträge in den GO-Termen aller Proteine aus einem vorherigen Experiment zum Interaktom von STAT3 am Helmholtz-Zentrum

für Umweltforschung erstellt. Alle drei Listen werden aus dem zweiten Tabellenreiter in dem Excel-Template „Fequently detected proteins“ abgerufen. Somit ist es möglich diese Listen zu aktualisieren. Weitere Filterelemente beziehen sich auf die Anzahl an verwendbaren Replikaten sowie einem Schwellwert für den T-Test und das Verhältnis. Aus diesen gesetzten Filtern und den zugehörigen Schwellwerten leiten sich die potentiellen Interaktionspartner ab, die bei der Messung bestimmt worden sind. Abschließend wird die unter Punkt 1.3.4 erwähnte Liste an Interaktionspartnern von der GeneCards Datenbank mit den potentiellen Interaktionspartnern abgeglichen.

3.4 Netzwerk erstellen

Als Grundlage für die Fitnessfunktion dient ein Netzwerk, das zentrisch zu dem zu untersuchenden Protein bzw. zu den zu untersuchenden Proteinen aufgebaut ist. Im Folgenden wird dieses Protein als Startprotein bzw. werden diese Proteine als Startproteine bezeichnet Das Netzwerk besteht dabei aus Kanten, den Proteininteraktionen und Knoten, den einzelnen Proteinen. Ausgehend von dem Startprotein kann das Netzwerk in Ebenen aufgegliedert werden. Die Startproteine werden dabei als Interaktionspartner der 0. Ebene oder Proteine der 0. Ebene definiert. Die Proteine, die direkte Interaktionspartner des Startproteins sind, werden im Folgenden als Interaktionspartner der 1. Ebene oder Proteine der 1. Ebene bezeichnet. Entsprechend sind Interaktionspartner der 2. Ebene Proteine, die ihrerseits direkte Interaktionspartner zu den Proteinen der 1. Ebene sind – sofern diese nicht schon in einer vorherigen Ebene vorgekommen sind. Die Ebene eines Proteins repräsentiert demnach die Pfadlänge zu dem Startprotein bzw. den Startproteinen.

Die Grundlage für das Netzwerk bildeten die Datenbanken MINT und IntAct. Von beiden Datenbanken wurde die jeweilige MITAB2.5 [Kerrien et al., 2007] - Datei heruntergeladen, die jeweils die gesamte Datenbank enthält. Beide Datenbanken befanden sich auf dem Stand vom 01.08.2012. Diese Dateien wurden nacheinander eingelesen und die entsprechenden Interaktionen wurden dem Netzwerk hinzugefügt.

Für die Erstellung der Netzwerke wurden ausschließlich Interaktionen gewählt, die zwischen menschlichen Proteinen gebildet werden. Zusätzlich wurden nur Interaktionen ausgewählt, die in der MINT-Datenbank einen Wert von 0,35 oder besser bzw. die in der IntAct-Datenbank einen Wert von 0,35 oder besser aufweisen konnten.

Von dem Startprotein wurden die direkten Interaktionspartner aus den ausgewählten Datenbanken herausgesucht. Diese Interaktionen wurden anschließend zu dem Netzwerk hinzugefügt, sofern

diese Interaktionen noch nicht im Netzwerk vorhanden waren. Von diesen direkten Interaktionspartnern wurden wieder alle direkten Interaktionspartner herausgesucht und dem Netzwerk hinzugefügt, sofern diese Interaktionen ebenfalls noch nicht im Netzwerk vorhanden waren. Dieser gerade beschriebene Schritt wurde so lange wiederholt, bis keine weiteren Proteine aus den Datenbanken zu dem Netzwerk hinzugefügt werden konnten.

3.5 Optimierung

In dieser Arbeit wurde eine Kombination aus der Evolutionären Strategie und dem Genetischen Algorithmus angewandt.

Zur Ermittlung des besten Parametersatzes wurde eine $[2+4(8+16)^{20}]^{100}$ – ES genutzt. Initialisiert wurden demnach zwei Startpopulationen mit je acht Individuen. Für jedes Individuum waren der Startwert des Verhältnisses (engl. *ratio*) 1 und der Startwert des p-Wertes 0,01. Aus den beiden Startpopulationen wurde viermal Eine kopiert, wobei die Auswahlwahrscheinlichkeit beider Populationen gleich war. Diese vier Kopien der Startpopulationen wurden in einer (8+16)-ES 20 Generationen lang verändert. Im Anschluss wurde die Fitness einer Population aus dem Durchschnitt der Fitnesswerte aller Individuen der betrachteten Population bestimmt. Aus den zwei Ausgangspopulationen und den vier neu erzeugten Populationen wurden die besten beiden Populationen ausgewählt. Diese bilden die Startpopulationen für den nächsten Durchlauf. Dieser Vorgang terminiert nach 100 Wiederholungen und das Individuum mit dem besten Fitnesswert repräsentierte den optimalen Parametersatz.

In der (8+16) – ES wurde aus dem Pool an acht Eltern-Individuen 16 mal eines kopiert, sodass 16 Kinder-Individuen entstanden. Im Anschluss wurden diese Kinder stochastisch, paarweise ausgewählt. Mit einer Wahrscheinlichkeit von 90 % wurden beide Individuen verändert. Jeder Parameter wurde dabei einzeln betrachtet. Zunächst wurden dabei die reellen Werte der Parameter in Bitstrings umgerechnet. Für eine hinreichend große Genauigkeit wurde die Länge dieses Strings auf 27 festgelegt. Anschließend wurden Mutation und Rekombination simuliert. Die Mutationswahrscheinlichkeit wurde auf 10 % festgesetzt. Das bedeutet, dass jede einzelne Position in einem Bitstring eine 10 %-ige Chance hatte, inkrementiert zu werden. Danach wurden die beiden mutierten Bitstrings mit einer Wahrscheinlichkeit von 25 % rekombiniert. Dabei wurde ein Ein-Punkt-Crossover durchgeführt, wobei dieser Punkt auf dem Bitstring stochastisch ermittelt wurde. Nachdem die 16 Kinder verändert wurden, wurde die Fitness aller Individuen bestimmt. Aus diesem Pool an 16 Kinder und Acht Eltern wurden die Acht Individuen mit der besten Fitness ausgewählt und der Vorgang beginnt von vorn. Nach der 20. Generation ist dieser Teilabschnitt beendet. Die gesamte $[2+4(8+16)^{20}]^{100}$ – ES ist in nachstehender Abbildung 2 visualisiert.

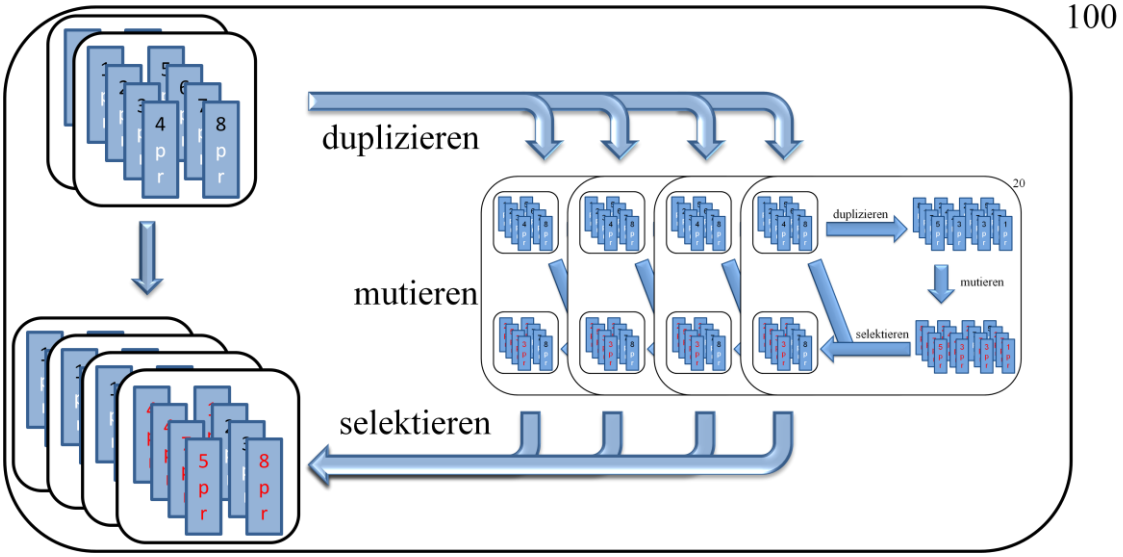


Abbildung 2: Visualisierung der $[2+4(8+16)^{20}]^{100}$ – ES
Hierbei sind links oben die beiden Populationen dargestellt, von denen viermal eine kopiert wird. Jede kopierte Population wird mit der (8+16)-ES, die auf der rechten Seite dargestellt sind, 20 Generationen lang verändert. Ein Individuum wird von einem blauen Rechteck mit einer schwarzen Nummer dargestellt. Die beiden Parameter der Optimierung sind mit „p“ und „r“ für Verhältnis und p-Wert beschrieben. Wurde in diesem Beispiel ein Individuum mutiert, sind die Nummer und die Parameter rot dargestellt.

3.5.1 Fitnessfunktion

Für eine Optimierung der Parameter wurden alle gemessenen, potentiellen Bindungspartner untersucht. Dazu wurde die Menge der je nach gewähltem Parametersatz verbleibenden Bindungspartner bewertet. Für diese Bewertung wurde eine Fitnessfunktion erstellt, die sich aus drei Termen zusammensetzt: Der Vollständigkeitsterm, der Pfadlängenterm und der Genauigkeitsterm. Als Testdatensatz zum erstellen dieser Funktion wurde ein Versuch zugrunde gelegt, der unter 3.2.1 beschrieben ist.

Vollständigkeitsterm

Dieser Teil der Funktion bezieht sich auf die Menge der potentiellen Bindungspartner, die – gemäß den gesetzten Parametern – aus allen in der Messung bestimmten Proteinen ausgewählt wurden. Im Testdatensatz waren rund 15 % der gemessenen Proteine Bindungspartner. In diesem Bereich erreicht dieser Term sein Optimum. Sind weniger Proteine ausgewählt, könnten einige Bindungspartner in der Liste fehlen. Sind mehr Proteine ausgewählt, könnten zu viele Proteine als Bindungspartner deklariert sein. Die Funktion ist in Abbildung 3 dargestellt.

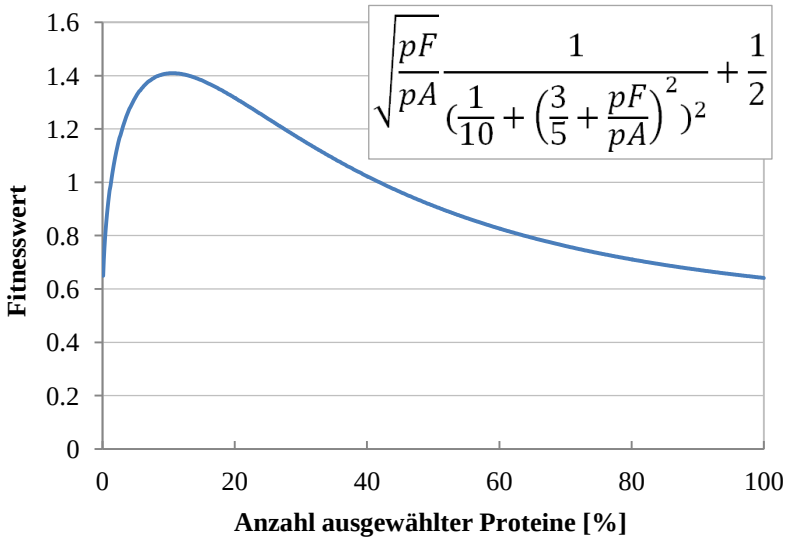


Abbildung 3: Vollständigkeitsterm der Fitnessfunktion
Der Verlauf sowie die entsprechende Formel des Vollständigkeitsterms sind aufgeführt. Dabei bezeichnet *pF* die, gemäß den gesetzten Parametern erhaltene Anzahl an potentiellen Bindungspartnern und *pA* bezeichnet die Anzahl aller in der Messung gefundenen Proteine.

Pfadlängenterm

Dieser Teil der Funktion bestimmt den mittleren Abstand der, gemäß den gesetzten Parametern verbleibenden, potentiellen Bindungspartner zu dem zu untersuchenden Protein bzw. zu den zu untersuchenden Proteinen. Als Grundlage wurde ein Netzwerk genutzt, dass zentrisch zu dem zu untersuchenden Protein, bzw. zentrisch zu den zu untersuchenden Proteinen ist. Weitere Details zum Aufbau und erstellen der Netzwerke sind in 3.4 beschrieben. Für jedes Protein, das potentieller Bindungspartner ist und sich im Netzwerk befindet, wurde die kürzeste Pfadlänge zum zu untersuchenden Protein im Netzwerk bestimmt. Gab es mehr als ein zu untersuchendes Protein, wurde der kürzere der Pfade ausgewählt. Für direkte Interaktionspartner galt die Pfadlänge 1. Die Pfadlänge wurde für jedes weitere Protein um eins erhöht, das zwischen den beiden Proteinen liegt, für die der kürzeste Pfad im Netzwerk bestimmt werden sollte. Zur Bestimmung der Pfadlänge zwischen zwei Proteinen im Netzwerk wurde der Bellmann-Fort Algorithmus [Cormen et al.; 2007] implementiert, der in nachstehender Abbildung 4 dargestellt ist. In diese Berechnung wurden nur Proteine einbezogen, die im Netzwerk enthalten waren.

```
01  für jedes v aus V
02      Distanz(v) := unendlich, Vorgänger(v) := keiner
03  Distanz(s) := 0

04  wiederhole n - 1 mal
05      für jedes (u,v) aus E
06          wenn Distanz(u) + Gewicht(u,v) < Distanz(v)
07          dann
08              Distanz(v) := Distanz(u) + Gewicht(u,v)
09              Vorgänger(v) := u

10  Ausgabe Distanz
```

Abbildung 4: Bellmann-Fort Algorithmus
Das Protein-Protein-Interaktionsnetzwerk kann als Graph aus Knoten und Kanten angesehen werden. Die Knoten stellen dabei Proteine und die Kanten Proteininteraktionen dar. Die Menge aller Knoten wird mit V beschrieben und ein Knoten aus V wird mit v oder u gekennzeichnet, wobei s der Startknoten ist. Die Menge aller Kanten ist E, wobei (u, v) eine Kante darstellt.

Anschließend wurde die Pfadlänge auf das Intervall [1;-1] normiert, um unterschiedlich große Netzwerke in der Fitnessfunktion gleich behandeln zu können. Dazu wurden alle Proteine der IntAct Datenbank und MINT Datenbank bestimmt, die an Interaktionen beteiligt waren, deren Schwellwert $\geq 0{,}35$ für die IntAct sowie $\geq 0{,}35$ für die MINT war. Weiterhin wurden nur Interaktionen mit Proteinen des menschlichen Organismus einbezogen. Aus diesen Proteinen wurden 1000 stochastisch ausgewählt und die mittlere Pfadlänge dieser 1000 Proteine zu den zu untersuchenden Proteinen im Netzwerk bestimmt. Dieser Vorgang wurde 100-mal wiederholt und der Durchschnitt gebildet, sodass die Pfadlänge von einer zufällig bestimmten Menge an Protein zu den zu untersuchenden Proteinen im Netzwerk bestimmt werden konnte. Die normierte Pfadlänge entspricht 0, wenn eine Menge an potentiellen Bindungspartnern eine mittlere Pfadlänge aufweist, die der Pfadlänge von einer zufällig bestimmten Menge an Protein gleicht. Wenn eine Menge an potentiellen Bindungspartnern eine mittlere Pfadlänge von 1 aufweist, entspricht die normierte Pfadlänge ebenfalls 1. Entspricht die mittlere Pfadlänge von einer Menge an potentiellen Bindungspartnern der maximal im Netzwerk möglichen Pfadlänge, wird die normierte Pfadlänge auf -1 gesetzt. Dies ist in nachstehender Abbildung 5 veranschaulicht.

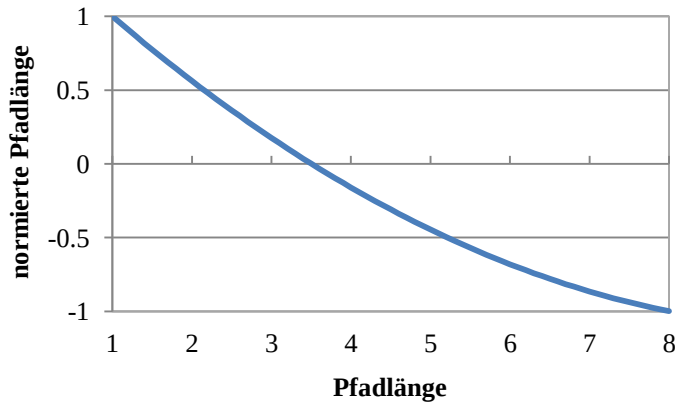


Abbildung 5: normierte Pfadlänge
Die normierte Pfadlänge ist eine Funktion durch die drei Punkte (1; 1), (mittlere Pfadlänge; 0) und (maximale Pfadlänge; -1). In dieser hier aufgeführten Funktion sind die mittlere Pfadlänge 3,5 und die maximale Pfadlänge 8.

Die normierte Pfadlänge wurde anschließend derart gewichtet, dass sich ähnliche Pfadlängen[8] deutlicher voneinander unterscheiden, wenn sie näher an der Pfadlänge zufällig ausgewählter Proteine sind. Somit wird erreicht, dass kleinere Verbesserungen der Pfadlänge8 weniger stark begünstigt werden, wenn sie nahe der maximal oder minimal möglichen Pfadlänge sind. Diese Wichtung wurde erreicht, indem die 3. Wurzel der normalisierten Pfadlänge gebildet wurde. Visualisiert ist der Pfadlängenterm in Abbildung 6.

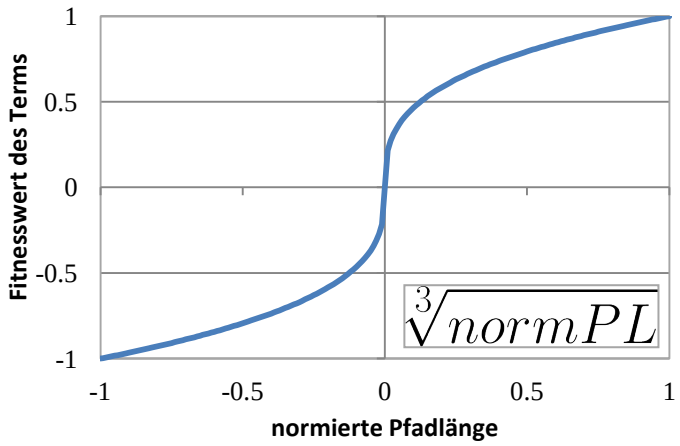


Abbildung 6: Pfadlängenterm der Fitnessfunktion
Der Verlauf sowie die entsprechende Formel des Pfadlängenterms der Fitnessfunktion sind aufgeführt. Hierbei wurde die normalisierte Pfadlänge von der Menge an potentiellen Bindungspartnern, die gemäß den gesetzten Parametern ausgewählt wurden, bestimmt. Die normierte Pfadlänge ist dabei mit *normPL* gekennzeichnet.

[8] Die mittlere Pfadlänge der gemäß gesetzten Parametern verbleibenden, potentiellen Bindungspartner.

Genauigkeitsterm

Als Referenz für die Genauigkeit dienten die bereits bekannten Bindungspartner, die in der GeneCards Datenbank [Safran et al., 2010; Shklar et al., 2005] aufgeführt sind. Dazu wurde die Liste aller Interaktionspartner, die nach der GeneCards Datenbank mit dem zu untersuchenden Protein interagieren, heruntergeladen.

Dieser Term bewertet eine Menge an potentiellen Bindungspartnern, die gemäß den gesetzten Parametern ausgewählt wurden, anhand der enthaltenen Bindungspartner, die ebenfalls in der GeneCards Datenbank stehen. Von diesen Bindungspartnern wurde ebenfalls die normalisierte Pfadlänge zu dem zu untersuchenden Protein, bzw. zu den zu untersuchenden Proteinen, berechnet. Die normierte Pfadlänge wurde auch hier anschließend derart gewichtet, dass sich ähnliche Pfadlängen[9] deutlicher voneinander Unterscheiden, wenn sie näher an der Pfadlänge zufällig ausgewählter Proteine sind. Somit wird erreicht, dass kleinere Verbesserungen der Pfadlänge8 weniger stark begünstigt werden, wenn sie nahe der maximal oder minimal möglichen Pfadlänge sind. Diese Wichtung wurde erreicht, indem die 3. Wurzel der normalisierten Pfadlänge gebildet wurde. In nachstehender Abbildung 7 ist dieser Term aufgeführt.

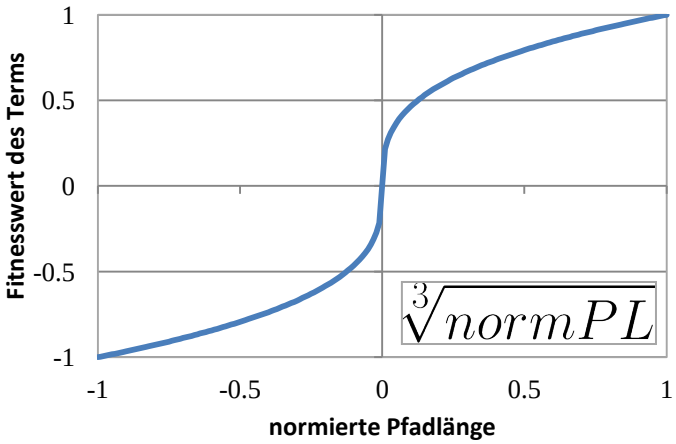


Abbildung 7: Genauigkeitsterm der Fitnessfunktion
Der Verlauf sowie die entsprechende Formel des Genauigkeitsterms der Fitnessfunktion sind aufgeführt. Hierbei wurde die normalisierte Pfadlänge von der Menge an potentiellen Bindungspartnern, die gemäß den gesetzten Parametern ausgewählt wurden und die ebenfalls in der GeneCards Datenbank annotiert sind, bestimmt. Die normierte Pfadlänge ist dabei mit *normPL* gekennzeichnet.

Diese drei Terme – Pfadlängenterm, Genauigkeitsterm und Vollständigkeitsterm – wurden zu nachstehender Formel zusammengefasst:

[9] Die mittlere Pfadlänge der gemäß gesetzter Parameter verbleibenden, potentiellen Bindungspartner.

$$\left(\left(\sqrt{\frac{pF}{pA}} \frac{1}{\left(\frac{1}{10} + \left(\frac{3}{5} + \frac{pF}{pA} \right)^2 \right)^2} + \frac{1}{2} \right) * \sqrt[3]{normPLg} \right) + \frac{\sqrt[3]{normPLp}}{2}$$

Abbildung 8: Fitnessfunktion
Die Formel berechnet den Fitnesswert eines Parametersatzes, der eine Menge an potentiellen Bindungspartnern aus allen gemessenen Proteinen bestimmt. Dabei ist *pF* die Anzahl an potentiellen Bindungspartnern und *pA* ist die Anzahl an allen gemessenen Proteinen. Die normalisierte Pfadlänge – berechnet aus der Menge an potentiellen Bindungspartnern – ist mit *normPLp* gekennzeichnet. Die normalisierte Pfadlänge – berechnet aus der Menge an potentiellen Bindungspartnern, die ebenfalls in der GeneCards Liste der Interagierenden Proteine mit dem zu untersuchenden Protein aufgeführt sind – ist mit *normPLg* gekennzeichnet.

Der Genauigkeitsterm und der Vollständigkeitsterm werden multipliziert, da sie ein gegenläufiges Verhalten zeigen. Während der Genauigkeitsterm für sich allein stehend nur einen Parametersatz bestimmt, der nur bekannte Bindungspartner hervorbringt, bestimmt der Vollständigkeitsterm für sich allein stehend nur einen Parametersatz der sehr viele, unter Umständen auch falsche Bindungspartner hervorbringt. Der Pfadlängenterm aus allen Bindungspartnern, die gemäß den gesetzten Parametern ausgewählt wurden, ermöglicht eine Feinabstimmung der Parameter.

3.6 Komplementäre Datentypen

Als Grundlage für die Anwendung von Dreieck-Netzwerk-Motiven und komplementären Daten dient die Bestimmung der gemeinsamen Interaktionspartner von zwei Proteinen. Diese beiden Proteine werden im Folgenden als potentiell interagierendes Proteinpaar bezeichnet. Die gemeinsamen Interaktionspartner des potentiell interagierenden Proteinpaares in einem PPIN werden als Nachbarn der zweiten Stufe bezeichnet. Zur Bestimmung dieser Nachbarn wurde der HIERDENC Algorithmus [Andreopoulos et al., 2007b] genutzt. Aus diesen Nachbarn und dem potentiell interagierendem Proteinpaar wurde genau dann ein Dreieck-Netzwerk-Motiv gebildet, wenn potentiell interagierendem Proteinpaar mindestens zwei gemeinsame Interaktionspartner (Nachbarn der zweiten Stufe) hat und mittels komplementären Daten eine Verbindung zwischen den potentiell interagierenden Proteinen bestimmt werden konnte. [Andreopoulos et al., 2007a] Ein Dreieck-Netzwerk-Motiv ist in Abbildung 9 visualisiert. In dieser Arbeit wurden die GO-Terme, strukturelle Domänen-Domänen Interaktionen und Literatur Kookkurrenzen als komplementäre Daten verwendet, die nachfolgend näher erläutert werden. In dieser Arbeit wurden die drei komplementären Datensätze sowie das zugrunde liegende Netzwerk erstellt. Die Zusammenführung dieser Daten

und die beschriebene Anwendung der Dreieck-Netzwerk-Motive auf das PPIN wurde von Bill Andreopoulos durchgeführt.

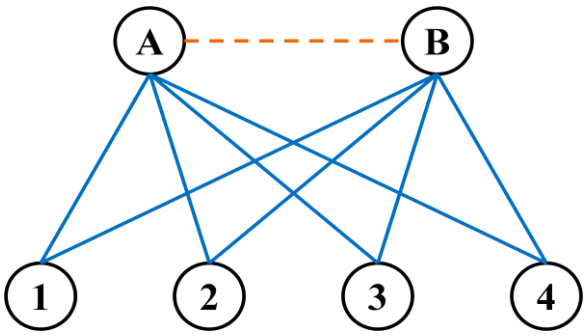


Abbildung 9: Dreieck-Netzwerk-Motiv
Die potentiell interagierenden Proteine sind mit A und B verdeutlicht. Beide haben vier Proteine (1, 2, 3 und 4) als gemeinsame Interaktionspartner. Hierbei sind Protein-Protein Interaktionen Blau und eine Verbindung durch komplementäre Daten ist Orange dargestellt.

3.6.1 GO-Terms

Der erste komplementäre Datensatz wurde aus den GO-Termen [Ashburner et al., 2000; Gene Ontology Consortium, 2011] erstellt. Dafür wurden alle GO-Terme von jedem Protein, das im Netzwerk enthalten war, heruntergeladen. Zu diesem Zweck wurde der AmiGO Webservice, der vom GO Consortium entwickelt worden ist [Carbon et al., 2009], genutzt. Für die Suchanfrage auf diesem Webservice wurden die UniProtKB-AN der einzelnen Proteine genutzt, sodass sicher gestellt werden konnte, dass nur GO-Terme der menschlichen Proteine herausgesucht wurden. Die einzelnen GO-Terme wurden dabei für jedes Protein separat und zusätzlich getrennt nach den jeweiligen Kategorien molekulare Funktion, biologischer Prozess und zelluläre Komponente gespeichert. Somit sind für jedes Protein, sofern die jeweiligen GO-Term Daten vorhanden waren, drei kleine Listen entstanden. Anschließend wurden alle Proteine paarweise miteinander verglichen. Besagte Trennung in die drei Kategorien erfolgte, weil alle Proteine zuerst hinsichtlich ihrer Lokalität (Kategorie: zelluläre Komponente) in den Zellen miteinander verglichen wurden. Für den Vergleich zweier Proteine durch deren GO-Terme wurde der G-SESAME Webservice genutzt [Du et al., 2009], mit dem es möglich ist zwei Listen an GO-Termen miteinander semantisch zu vergleichen. Die Eingabe beider GO-Term Listen beider Proteine erfolgte mit den voreingestellten Parametern ("is_a" Beziehung 0,8; "part_of" Beziehung 0,6). Zwischen beiden wurde die semantische Ähnlichkeit, die im Intervall [0,1] lag, ermittelt. Als Schwellwert wurde 0,8 als stringent genug bestimmt. Als Richtwert diente der Wert, der zwischen der Semantischen Go-Term Analyse von STAT1 und STAT3 durch den G-SESAME Webservice ermittelt wurde. Für alle Proteinpaare, deren semantische Ähnlichkeit bezüglich ihrer Lokalität in den Zellen über diesem Schwellwert lag, wurde zusätzlich geprüft, wie hoch die semantische Ähnlichkeit (ebenfalls bestimmt mittels

GO-Term Vergleich) bezüglich der molekularen Funktion bzw. des biologischen Prozesses war. Dafür wurde ebenfalls ein Schwellwert von 0,8 festgelegt und als stringent genug erachtet. Als Richtwerte dienten hierbei ebenfalls die Semantischen GO-Term Analysen von STAT1 und STAT3. Ein Proteinpaar wurde demnach dem komplementären Datensatz hinzugefügt, wenn die semantische Ähnlichkeit der zellulären Lokalität über 0,8 und entweder die semantische Ähnlichkeit der molekularen Funktion oder die semantische Ähnlichkeit des biologischen Prozesses ebenfalls über 0,8 lag.

3.6.2 Strukturelle Domain-Domain Interaktionen

Als Grundlage zur Bestimmung der Domänen-Domänen Interaktionen diente die SCOPPI Datenbank. Im ersten Schritt wurden alle Domänensequenzen, die auf dieser Datenbank hinterlegt waren, heruntergeladen (Stand: 29.08.2012) und im FASTA-Format gesichert. Zusätzlich wurde vermerkt, welche Domänen miteinander interagierten. Im zweiten Schritt wurden alle Proteinsequenzen, der im Netzwerk vorhanden Proteine, ebenfalls heruntergeladen und im FASTA-Format gespeichert. Um Abschließend zu bestimmen, welche Proteine aufgrund von Domänen-Domänen Interaktionen in den komplementären Datensatz hinzugefügt werden konnten, wurde eine aktuelle Version des BLAST standalone installiert. Die Installation und alle weiteren Schritte wurden angelehnt an die BLAST Hilfe [Camacho et al., 2008] durchgeführt. So wurde zuerst aus den FASTA-Dateien, die aus den SCOPPI Daten erzeugt wurden, eine BLAST Datenbank erstellt. Gegen diese Datenbank wurden anschließend alle FASTA-Dateien, die aus den Proteinen des Netzwerkes erzeugt wurden, gesucht. Die Ausgabe des BLAST Laufes wurde dahingehend angepasst, dass sowohl der e-Wert, die Sequenzidentität, Gesamtlänge des Treffers sowie Startindex und Endindex der Suchsequenz des jeweiligen Treffers angegeben wurden. Als Treffer werden die Domänensequenzen bezeichnet, die aus der SCOPPI entnommen wurden und als Suchsequenzen sind die Sequenzen der Proteine bezeichnet, die aus dem Netzwerk entnommen wurden. Bei der Zuordnung der Domänen zu den Proteinsequenzen wurden nur Treffer beachtet, die einen e-Wert $\leq 0{,}01$ und eine Sequenzidentität von ≥ 30 % aufweisen konnten. Zusätzlich dazu mussten mindestens 75 % der Domänensequenz in der Proteinsequenz auftauchen. Diese Schwellwerte wurden angelehnt an [Andreopoulos et al., 2009], bestimmt. Abschließend wurde anhand der miteinander interagierenden Domänen festgestellt, welche Proteinsequenzen miteinander interagieren können. Diese Proteinpaare wurden diesem komplementären Datensatz hinzugefügt.

3.6.3 Literatur Kookkurrenz

Zur Ermittlung der Literatur-Kookkurrenzen wurde die semantische Suchmaschine GoPubMed [Doms & Schroeder, 2005] genutzt. Dazu wurden von allen Proteinen, die in der Datenbank enthalten waren, Suchanfragen an GoPubMed gestellt, um festzustellen, in wie vielen wissenschaftlichen Veröffentlichungen sie enthalten waren. Anschließend wurden von allen Proteinen paarweise Anfragen gestellt, um festzustellen, wie viele Publikationen mit beiden Proteinen veröffentlich waren. Bei beiden Anfragen wurden die Gennamen der Proteine verwendet und der Zeitraum, in dem die Anfragen gestellt wurden belief sich auf August und September 2012. Für die Umsetzung dieser großen Menge an Suchanfragen wurde Rücksprache mit der Transinsight GmbH gehalten, die die semantische Suchmaschine verwaltet. Im Anschluss wurde eine Version der Blosum Kookkurrenz Bewertung genutzt, um festzustellen, ob zwei Proteine p_1 und p_2 in den Dokumenten von PubMed häufig genug nebeneinander auftreten; dargestellt in nachstehender Gleichung:

$$log \frac{Prob(p_1\ und\ p_2)}{Prob(p_1) * Prob(p_2)} > 10$$

Wobei *Prob(p_1)* die Wahrscheinlichkeit[10] ist, mit der das Protein p_1 ausgewählt werden kann. Die Gesamtanzahl der Dokumente in GoPubMed belief sich auf 22.231.158 in besagtem Zeitraum. Der Schwellwert von 10 wurde angelehnt an [Andreopoulos et al., 2009] ausgewählt. Wurde für ein Proteinpaar ein Wert über 10 ermittelt, wurde das entsprechende Proteinpaar dem komplementären Datensatz hinzugefügt.

[10] In diesem Falle die Anzahl der Literaturstellen in denen das Protein auftritt, geteilt durch die Gesamtanzahl an Literaturstellen in der Datenbank.

4 Entwickeltes Softwaretool

Zur Automatisierung der Datenverarbeitung und Vereinfachung der Handhabung wurde eine grafische Benutzeroberfläche (GUI; engl. *grafical user interface*) implementiert. Die Toolbox enthält einen „data“, einen „resources“ und einen „serialisedPPIs“ Ordner. Zusätzlich enthält diese eine ausführbare JAR-Datei, mit der das Programm gestartet werden kann. Bei jeder Anwendung, die Daten erstellt, wird der „data“ Ordner als Speicherort vorgeschlagen. Im „resources“ Ordner sind die Template-Dateien, deren CSV-Dateien sowie die Ordner „jarFiles“ und „preExpSerPIPs“ hinterlegt. Der „jarFiles“ Ordner enthält alle zusätzlich für das Programm benötigte Java-Klassen. Im „preExpSerPIPs“ Ordner sind temporäre, serialisierte Dateien abgelegt, die aus den CSV-Dateien erzeugt werden. Der Ordner „serialisedPPIs“ wird als Speicherort für neue PPINs vorgeschlagen und enthält bereits einige serialisierte PPIN, die durch die Dateiendung „.ppin“ erkennbar sind.

4.1 Parsen der MS-Daten

Triple Silac Daten, die mittels Proteom Discoverer ausgewertet wurden, können mit Hilfe des Parsers schnell in eine übersichtliche Form gebracht werden. Die Ausgabe des Proteom Discoverer muss dahingehend angepasst werden, dass nur die oberste Ebene der Ergebnisse in eine Datei exportiert wird. Triple Silac Daten, die mittels MaxQuant analysiert worden sind, können ebenfalls geparst werden, indem die *proteinGroups*-Datei eingelesen wird. Zusätzlich muss die Liste an interagierenden Proteinen mit dem zu untersuchenden Protein von der GeneCards Datenbank in einer Datei gesichert werden; siehe dazu auch 1.3.4. Wichtig sind dabei nur die ersten beiden Spalten der Liste, die im Folgenden als GeneCard-Liste bezeichnet wird. Beim kopieren der GeneCard-Liste aus dem Browser werden die Fußnoten direkt mit als Text kopiert – der interne Parser erkennt dennoch die UniProtKB-ANs. Zur übersichtlichen Darstellung der Daten wurde ein Excel-Template entworfen, das an die Ausgabedaten angepasst ist. Ein Template für sechs Replikate ist in Anhang A: Excel-Template abgebildet. Diese Templates sind aufgeteilt in vier Sektionen: die Proteininformationen sowie den Informationen über die H/L, M/L und H/M Ansätze. Für jeden der drei Ansätze sind die einzelnen normierten Verhältnisse jedes Replikates, statistische Informationen (Anzahl Replikate, in denen etwas gemessen wurde; normierter Mittelwert der Verhältnisse; Standardabweichung; p-Wert des T-Testes) und Filteroptionen angegeben**.** Zur Normierung wird der Mittelwert der Mediane der Verhältnisse berechnet, wodurch sich Mischfehler bei der Probenanfertigung korrigieren lassen. Vor der Berechnung der statistischen Werte wurden die normierten Verhältnisse

zur Basis 2 logarithmiert. In den Filterfunktionen besteht die Möglichkeit biotinylierte Proteine, eine zu geringe Anzahl an Replikaten und Chaperone oder Proteolyse betreibende Proteine heraus zu filtern. Zusätzlich kann der Schwellwert des Verhältnisses und der Schwellwert des p-Wertes des T-Testes geändert werden. Die Ausgabe des MS-Parsers erfolgt für alle drei verschiedenen Ansätze getrennt, sodass jeder Ansatz in einer eigenen Datei ausgegeben wird oder die Ausgabe erfolgt zusammen in einer Datei. Der Inhalt der Dateien wird anschließend in das vorgefertigte Template kopiert. Momentan ist es möglich Experimente mit drei oder sechs Replikaten zu parsen. Entsprechend gibt es für beide ein Excel-Template. Einen Überblick über die GUI Optionen für diesen Parser gibt Abbildung 10.

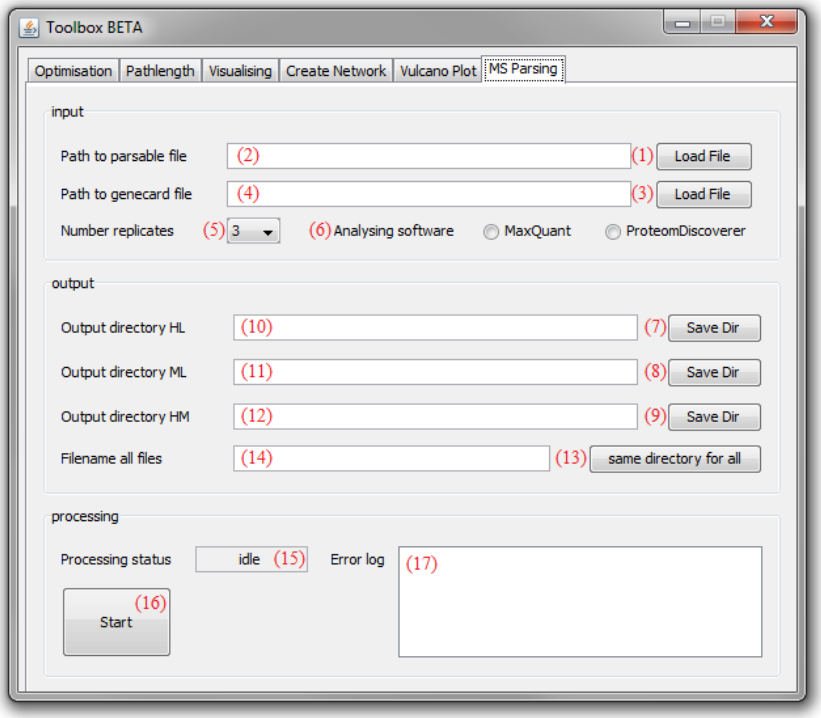


Abbildung 10: *MS-Parsing* Tab
Unter (1) kann die Datei geladen werden, die geparst werden soll (Ausgabedatei von Proteom Discoverer oder *proteinGroups*-Datei von MaxQuant). Unter (3) kann die Datei mit der GeneCards-Liste eingelesen werden. (2) und (4) stellen die Pfade zu beiden Dateien dar, die auch manuell geändert werden können. (5) ermöglicht eine Auswahl der Anzahl der Replikate und (6) bestimmt die Software, mit der die Daten ausgewertet wurden. Mit (7-9) können die Pfade für die einzelnen Ausgabedateien bestimmt werden. Diese Pfade werden unter (10-12) dargestellt, die ebenfalls manuell editiert werden können. Sollen alle Ansätze in einer Datei gespeichert werden, muss der Ausgabepfad aller Ansätze überein stimmen. Mit (13) kann dies in einem Schritt durchgeführt werden. Der unter (13) angegebene Pfad wird anschließend in (10-12) übernommen. In (14) kann ein beliebiger Name für diese Datei mit den gemeinsamen Daten angegeben werden. (15) beschreibt den aktuellen Status des Parsers und (16) startet den Parser. Sollte ein Fehler bei der Eingabe aufgetreten sein, wird dieser unter (17) im Error log angezeigt.

4.2 Darstellen der MS-Daten

Nachdem die Auswertung der MS-Daten erfolgte, ist es möglich diese in einem Vulcano Plot[11] für den ersten Überblick zu visualisieren. Vorbereitend muss das mit Daten versehene Excel-Template (beschrieben unter 4.1) in eine CSV-Datei (*comma-separated value*) überführt werden. Dazu in der geöffneten Excel-Datei „speichern unter“, „andere Formate“ und als Dateityp „CSV (Trennzeichen-getrennt) (*.csv)“ auswählen und abspeichern. Wichtig ist, dass während dem abspeichern in eine CSV-Datei keine Filter gesetzt sind, weil das zu einem fehlerhaften Plot führen kann. Der Vulcano Plot wird anhand der Daten in dieser gesicherten CSV-Datei erstellt. Dabei wird zwischen den verschiedenen Ansätzen, H/L, M/L und H/M unterschieden. Zusätzlich werden nur Proteine beachtet, die wenigstens zwei Replikate haben und zudem keine Chaperone, Proteolysefaktoren und Proteine sind, die eine Domäne haben, die an Biotin binden kann. Der Plot wird als Pixelgrafik erzeugt und als JPG-Datei gespeichert, wobei sich die Gesamtgröße auf 3840 x 2160 Bildpunkte beläuft. Die Achsenskalierung des Plots wird automatisch anhand der Daten, die dargestellt werden sollen, berechnet. Sie ist derart angepasst, dass die äußersten Punkte direkt an der Plotgrenze liegen. Der Betrag des Verhältnisses ist dabei auf beiden Seiten äquivalent, damit keine Verzerrung entsteht. Somit ist gewährleistet, dass der Plot die darzustellenden Daten optimal verteilt. Zur einheitlichen Darstellung mehrerer Plots mit unterschiedlichen Skalierungen der Daten (beispielsweise liegt der kleinste p-Wert des einen Plots bei 10^{-4}, der kleinste p-Wert des anderen Plots bei 10^{-8}) ist es möglich, die Skalierung beider Parameter jeweils auf eine festgelegte Grenze zu erweitern, obwohl dies anhand der Daten nicht nötig gewesen wäre. Sollten die darzustellenden Daten höhere Werte (bei dem Verhältnis) bzw. kleine Werte (bei dem p-Wert) enthalten, als durch die entsprechende Grenze festgelegt wurde, ist die jeweils festgelegte Grenze hinfällig. Der Plot selbst zeichnet graue Datenpunkte auf weißem Hintergrund. Zusätzlich ist es möglich eine beliebige Anzahl an Datenpunkten in einer der vorgegebenen Farbe darzustellen. Die Einfärbung der Datenpunkte erfolgt in der Reihenfolge, in der die Farben angegeben sind. Für den Fall der doppelten Farbcodevergabe an einen Datenpunkt, bleibt der erste Farbcode bestehen. Für die Auswahl eines Datenpunktes muss die entsprechende UniProtKB-AN eingetragen werden. Es ist weiterhin möglich alle Datenpunkte mit jeweils gleichem Farbcode mit einer Beschriftung zu versehen. Die Beschriftung erfolgt oberhalb des Datenpunktes als dessen Genname. Einen Überblick über die GUI Optionen für die Darstellung der AP/MS Daten mittels Vulcano Plot gibt Abbildung 11.

[11] Das durchschnittliche Verhältnis einer Messreihe wird gegen den p-Wert dieser Messreihe aufgetragen. Eine Messreihe bezeichnet in diesem Fall alle gemessenen Replikate von einem Peptid. Dabei ist das Verhältnis linear auf der Abszisse und der p-Wert logarithmisch auf der Ordinate skaliert.

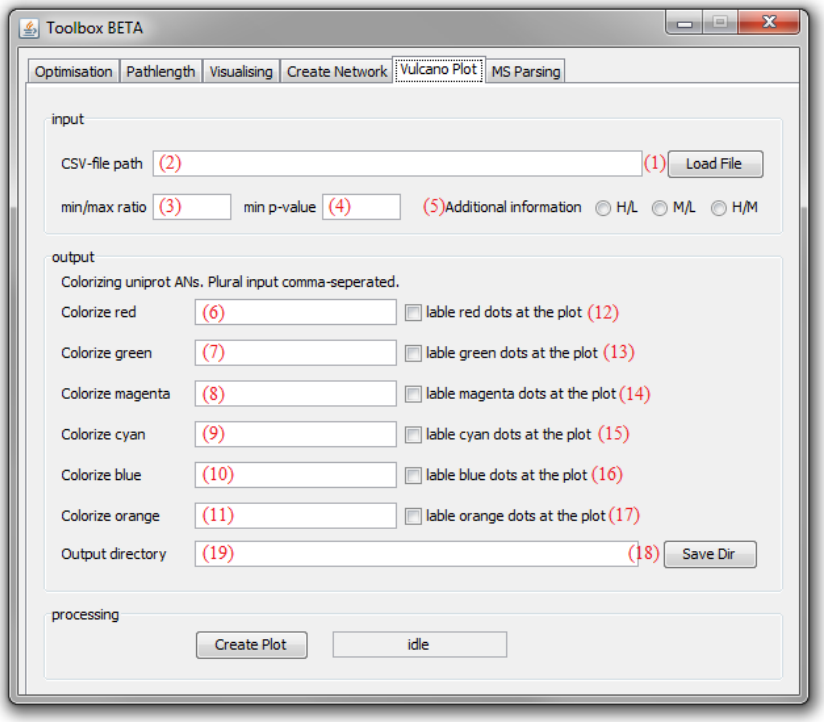


Abbildung 11: *Vulcano Plot* Tab der Toolbox
Unter (1) kann die CSV-Datei geladen werden, deren Daten dargestellt werden sollen. In (2) wird der Dateipfad dargestellt, der auch manuell geändert werden kann. Bei (3) und (4) können die Grenzen für eine Erweiterung der Skalierung für das Verhältnis (engl. *ratio*) bzw. den p-Wert eingetragen werden. Durch (5) ist es möglich den Ansatz, der dargestellt werden soll, auszuwählen. In (6-11) können die einzufärbenden Datenpunkte – Komma-separiert – eingetragen werden (UniProtKB-AN verwenden). Durch (13-17) kann der jeweiligen Menge an Datenpunkten eine Beschriftung hinzugefügt werden. Mit (18) kann der Pfade für die Ausgabedateien bestimmt werden. Dieser Pfad wird unter (19) dargestellt, der ebenfalls manuell editiert werden kann. (20) beschreibt den aktuellen Status der Anwendung und (21) startet die Anwendung.

4.3 Erstellen der Netzwerke

Ein PPIN kann, wie schon in 3.4 beschrieben, aus einer MITAB2.5-Datei oder einem bereits bestehenden Netzwerk erstellt werden. Als verwendbare Datenbanken stehen die IntAct und die MINT zur Verfügung für die jeweils ein Schwellwert für deren annotierte Interaktionen angegeben werden kann. Zusätzlich kann nach Interaktionen gefiltert werden, die ausschließlich zwischen menschlichen Proteinen stattfinden. Es besteht die Möglichkeit, dass das Netzwerk ausgehend von einem Startprotein, mehreren Startproteinen (jeweils zentrisch) oder gar keinen Startprotein (nicht zentrisch) erstellt wird. Bei letzterem werden alle existierenden Interaktionen der Quelle, die den bereits erwähnten Parametern entsprechen, zu dem Netzwerk hinzugefügt. Dies bietet sich bei sehr großen MITAB2.5-Dateien an. Damit ein Netzwerk für die Pfadlängenuntersuchungen weiter verwendet werden kann, muss es zentrisch aufgebaut sein. Somit ist es zu empfehlen, dass eine große MITAB2.5-Datei zuerst in ein nicht zentrisches PPIN geparst wird und anschließend aus diesem

nicht zentrischem PPIN ein zentrisches PPIN geparst wird. Während dem Parsen wird permanent ein Statusupdate geliefert, der den aktuellen Stand des Parsens veranschaulicht. Die Ausgabe erfolgt in eine serialisierte PPIN-Datei, die in 3.4 näher beschrieben ist. Einen Überblick über die GUI Optionen für den Netzwerkparser gibt nachstehende Abbildung 12.

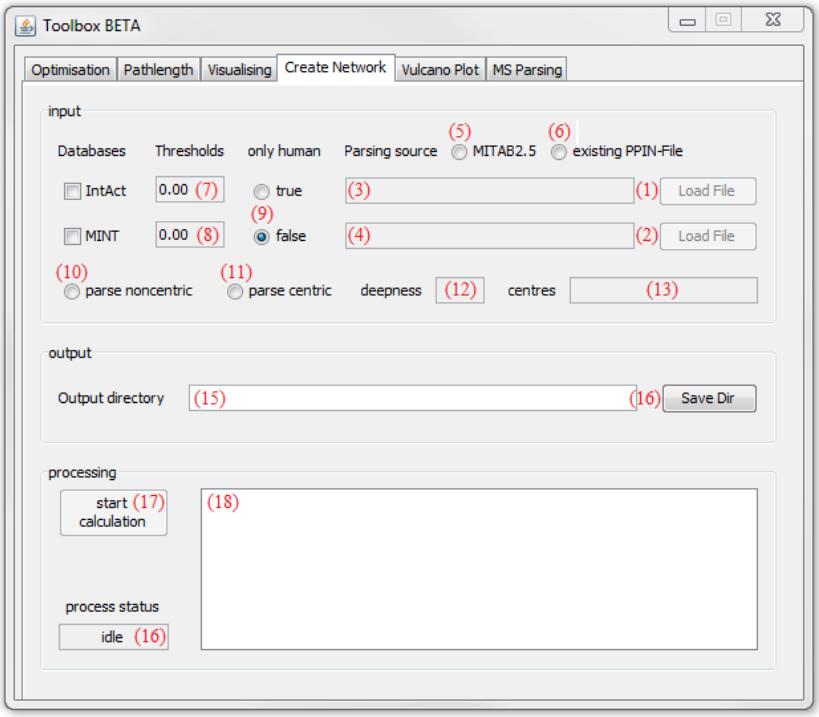


Abbildung 12: *Create Network* Tab der Toolbox
Durch (1) und (2) können die Datenbanken geladen werden und (3) und (4) stellen die Pfade zu beiden Dateien dar, die auch manuell geändert werden können. Mit (5) und (6) kann die Quelledatei für den Parser bestimmt werden. In (7) und (8) können die Schwellwerte (engl. *threshold*) für die IntAct bzw. die MINT gesetzt werden. Durch (9) ist es möglich, nur Interaktionen an denen menschliche Proteine beteiligt sind, zu beachten. Mit (10) werden alle Interaktionen der Quelldatei – gemäß den gesetzten Parametern – eingelesen (nicht zentrisch). Durch (11) kann das Netzwerk zentrisch geparst werden. Ist (11) ausgewählt, können in (12) die maximale Ebene und in (13) die Startproteine ausgewählt werden. Mit (14) kann der Pfad für die Ausgabedateien bestimmt werden. Dieser Pfad wird unter (15) dargestellt, der manuell editiert werden kann. (16) beschreibt den aktuellen Status der Anwendung und (17) startet die Anwendung. In (18) werden die Statusupdates ausgegeben.

4.4 Visualisieren der Daten des Netzwerkes

Die Daten aus einer PPIN-Datei können auf verschiedene Weise visualisiert werden. Zum einen ist es möglich eine Textausgabe aller Interaktionen oder aller Proteine des ausgewählten Netzwerkes zu erstellen. Zum Anderen kann das Netzwerk dynamisch visualisiert werden.

Einen Überblick über die GUI Optionen für die Visualisierung der Daten eines Netzwerkes gibt Abbildung 13.

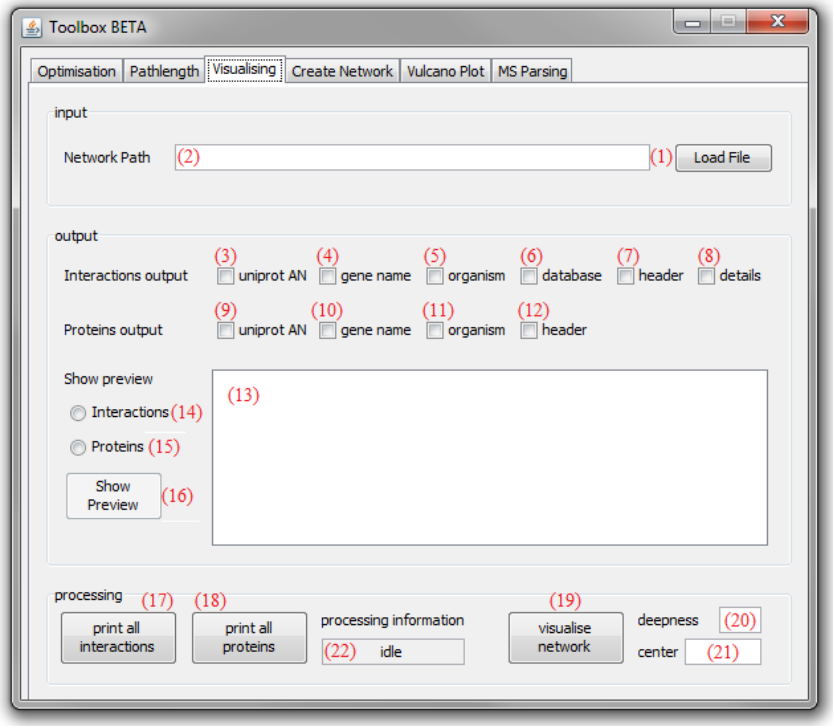


Abbildung 13: *Visualising* Tab der Toolbox
Unter (1) kann die PPIN-Datei geladen werden, deren Daten dargestellt werden sollen. In (2) wird der Dateipfad dargestellt, der auch manuell geändert werden kann. In (3-8) wird die Ausgabedatei der Interaktionen angepasst und in (9-12) wird die Ausgabedatei der Proteine angepasst. Dabei kann in (13) eine Vorschau des Ausgabeformates angeschaut werden, dass an einem Beispiel die Auswahl zeigt. Mit (14) kann die Vorschau für die Ausgabedatei der Interaktionen und mit (15) kann die Vorschau für die Ausgabedatei der Proteine angeschaut werden, wenn nach deren Auswahl (16) betätigt wird und (3-8) bzw. (9-12) entsprechend ausgewählt ist. Mit (17) kann die Ausgabe der Datei mit den Interaktionen und mit (18) kann die Ausgabe der Datei für die Proteine erfolgen. Durch (19) kann das Netzwerk dynamisch visualisiert werden, wobei in (20) der Genname des Proteins für das Zentrum des darzustellenden Netzwerkes eingetragen werden muss. In (21) muss die maximale Ebene, bis zu der visualisiert werden soll eingetragen werden. (22) beschreibt den aktuellen Status der entsprechenden Anwendung.

Für die Textausgabe der Interaktionen wird eine Textdatei erstellt, die, wenn gewünscht, Informationen zu allen UniProtKP-AN, allen Gennamen, allen Organismen und den Datenbanken der jeweiligen Proteine enthält. Dabei wird jede Interaktion auf einer Zeile dargestellt. Zusätzlich kann eine Kopfzeile (engl. *header*) eingefügt werden, die die dargestellten Informationen zusammenfasst. Weiterhin kann eine Fußzeile eingefügt werden, in der aufgeführt ist, wie viele Interaktionen aus den einzelnen Datenbanken entstammen. Die Textausgabe der Proteine beinhaltet ebenfalls die UniProtKB-AN, die Gennamen, die Organismen und eine Kopfzeile, die ebenfalls die dargestellten Informationen zusammenfasst. Jedes Protein wird in einer separaten Zeile dargestellt. Die dynamische Darstellung des Netzwerkes beläuft sich auf einen zentrisch ausgerichteten Graphen, der jeweils nur einen Knoten (bzw. ein Protein) im Zentrum visualisieren kann. (siehe z.B. Abbildung 20) Durch Auswahl eines Knotens der Darstellung wird das Netzwerk derart verschoben, dass der

ausgewählte Knoten, statt dem vorherigen, im Zentrum erscheint. In dieser Darstellung sind die Gennamen der Proteine angegeben.

4.5 Pfadlängenanalysen des Netzwerkes

Für die Fitnessfunktion ist es notwendig eine Pfadlängenanalyse durchzuführen. In einer derartigen Analyse werden die Pfadlängen von einer Liste an Proteinen zu dem Startprotein bzw. zu den Startproteinen eines ausgewählten Netzwerkes (siehe 3.4) bestimmt. Für diese Analyse werden zwei PPIN-Dateien geladen. Eine dieser beiden Dateien enthält das Netzwerk, in dem die Pfadlängenbestimmung durchgeführt wird. Aus der anderen PPIN-Datei wird die Proteinliste entnommen. Alle Proteine aus der einen PPIN-Datei werden somit im Netzwerk der anderen PPIN-Datei gesucht. Einen Überblick über die GUI-Optionen für die Pfadlängenanalyse eines Netzwerkes gibt Abbildung 14.

Die Suche selbst wird durch eine Form des Bellmann-Fort Algorithmus realisiert. Für die Suche von Proteinen in einem ausgewählten Netzwerk stehen zwei Suchoptionen zur Auswahl: Die Suche aller Proteine aus der Liste in dem Netzwerk oder die Suche einer festlegbaren Anzahl an Proteinen aus der Liste mit einer festlegbaren Anzahl an Iterationen. Dabei wird bei jeder Iteration die festgelegte Anzahl an Proteinen zufällig aus der Liste entnommen. Diese Option fand bei der zufälligen Pfadlängenbestimmung Verwendung, wobei 100 mal 1000 Proteine aus der Liste im Netzwerk gesucht wurden. Die Ausgabe erfolgt dabei in dem Textfenster. Darin wird in je einer eigenen Zeile angegeben: Wie viele Proteine nicht im Netzwerk gefunden wurden; wie viele Proteine der einzelnen Ebenen – beginnend mit der 0. Ebene – gefunden wurden; die durchschnittliche Pfadlänge sowie die drei durch ein Leerzeichen getrennten Terme der quadratischen Funktion der normierten Pfadlänge. Eine weitere Suchoption besteht darin, den Pfad zwischen zwei Proteinen in dem eingeladenen Netzwerk direkt zu bestimmen. Für diesen Vorgang muss von beiden Proteinen die UniProtKB-AN eingetragen werden. Die Ausgabe erfolgt in dem Textfenster und ist Beispielhaft in Abbildung 15 angeben.

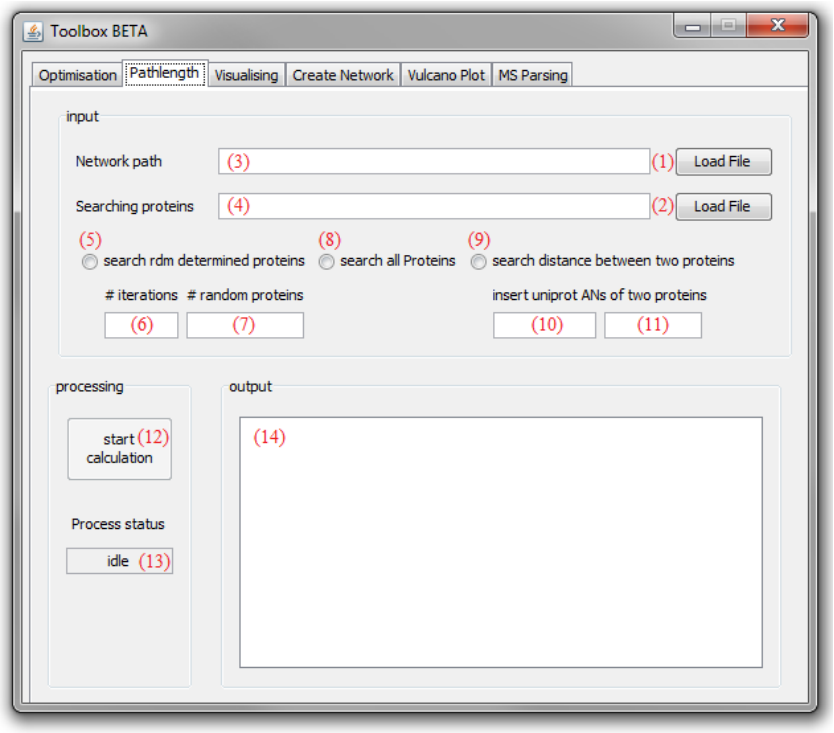


Abbildung 14: *Pathlength* Tab der Toolbox
Durch (1) und (2) können die Datenbanken geladen werden und (3) und (4) stellen die Pfade zu beiden Dateien dar, die auch manuell editiert werden können. Mit (5) kann die Zufallssuche in einer selbst bestimmten Anzahl an Iterationen (6) von einer selbst bestimmten Anzahl an Proteinen der Proteinliste – einzutragen in (7) – durchgeführt werden. Mit (8) werden alle Proteine der Proteinliste im Netzwerk gesucht. (9) ermöglicht die Bestimmung des Pfades zwischen zwei Proteinen, deren UniProtKP-ANs in (10) bzw. (11) einzutragen sind. In (12) wird die Ausgabe abgebildet, (13) beschreibt den aktuellen Status der Anwendung und (14) startet die Anwendung.

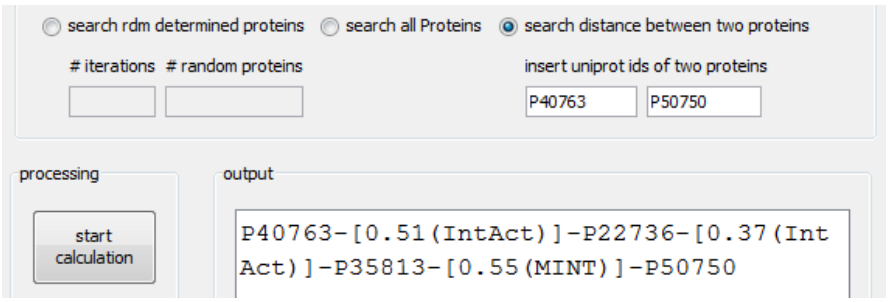


Abbildung 15: Beispiel-Suchergebnis des Pfades zwischen zwei Proteinen
Hierbei wurde der Pfad zwischen den Proteinen STAT3 (P40763) und CDK9 (P50750) ermittelt. Zwischen je zwei UniProtKB-ANs steht der Wert der Interaktion in eckigen Klammern. Diese Werte entsprechen den in den Datenbanken angegebenen Werten, wobei die Datenbank, aus der die Interaktion entnommen wurde in runden Klammern hinter dem Wert der Interaktion angegeben ist. Besitzt eine Interaktion Einträge in mehreren Datenbanken, so sind in den eckigen Klammern entsprechend mehr Datenbanken abgebildet. Besitzt eine Interaktion aus einer Datenbank keinen Wert, ist DNE statt einem Wert in den eckigen Klammern abgebildet.

4.6 Implementierung der Optimierung

Die Optimierung der beiden Parameter Verhältnis und p-Wert nimmt eine zentrale Stellung in diesem Programm ein. Der Algorithmus der Optimierung ist in 3.5 beschrieben. Die drei Ansätze H/L, M/L und H/M müssen dabei getrennt voneinander optimiert werden, wobei die Optimierung des H/M Ansatzes noch weiterer Anpassungen bedarf. Bei der Eingabe der sechs[12] für die Optimierung erforderlichen Parameter ist darauf zu achten, dass jede neue Population μ' auf einem eigenen Thread prozessiert wird. Somit ist es nicht empfehlenswert, wenn μ' die Anzahl an CPU-Kernen überschreitet. Für die Optimierung ist ein Netzwerk erforderlich, dass das zu untersuchende Protein im Zentrum enthält. Dies kann, wie unter 4.3 beschrieben, erstellt werden. Weiterhin wird die CSV-Datei von den Messergebnissen des zu untersuchenden Proteins benötigt. Diese Datei kann mit Hilfe der Anweisungen aus 4.1 erstellt werden. Zusätzlich sind die drei Terme der quadratischen Funktion der normalisierten Pfadlänge erforderlich. Mit Hilfe der Anweisungen von 4.5 können diese ermittelt werden, wobei die zuletzt in der Textbox (Abbildung 14, Nr.: 14) abgebildeten Terme automatisch in die Eingabemaske der quadratischen Funktion übernommen werden. Als Startparameter für die Optimierung sind 1,0 für das Verhältnis (engl. *ratio*) und 0,01 für den p-Wert voreingestellt. Das Programm benötigt zusätzlich die UniProtKB-AN des Proteins, für das die Optimierung durchgeführt werden soll. Als ersten Schritt muss nach Eingabe aller genannten Parameter der Optimierungsvorgang initialisiert werden. Damit wird ermittelt, wie viele Proteine in der Messung gefunden wurden, wie viele dieser Proteine in der GeneCards Datenbank vorkommen und welche mittleren Pfadlängen diese beiden Proteinmengen aufweisen. Anschließend kann die Optimierung selbst gestartet werden. Dies kann, in Abhängigkeit der gewählten Parameter einige Stunden in Anspruch nehmen. Der aktuelle Status des Prozesses ist in der Textbox dargestellt, wobei durch „overall try number: *a*/*m*“ die aktuelle Iteration *a* genannt und durch „inner try number: (*b*/λ') *c*/*n*“ mit *b* die entsprechende neue Population und mit *c* die aktuelle Iteration der Isolation genannt wird. Die Ergebnisse werden in einer Textdatei abgelegt, die schematisch in Abbildung 16 dargestellt ist. In dieser Datei sind alle während der Optimierung berechneten Fitnesswerte und alle UniProtKB-ANs der Proteine des besten Fitnesswertes aufgeführt, aufgetragen nach ihrem Abstand im Netzwerk zu dem Startprotein bzw. den Startproteinen. Einen Überblick über die GUI Optionen für die Optimierung gibt Abbildung 17.

[12] $[\mu'+\lambda'(\mu+\lambda)^n]^m$-ES mit der Anzahl der vorhandenen Populationen μ'; der Anzahl der neuen Populationen λ'; der Anzahl an Iterationen m; die Anzahl der Elternelemente μ; der Anzahl der Kinderelemente λ und der Anzahl an Iterationen der Isolation m.

fitness	ratio	pValue	normPL	all	not@Net	tier 0	tier 1	tier 2	tier 3	tier 4	tier 5	tier 6
1.6325	1.0648	0.1736	0.1413|1	82|3	12|0	1|1	1|1	9|1	49|0	8|0	1|0	1|0

not@Net	tier 0	tier 1	tier 2	tier 3	tier 4	tier 5	tier 6
O43615	P42224	P52630	P09936	O00273	O00233	O75396	P28838
P00374			P16930	O14929	O00762		
P09972			P27348	O43583	P00568		

Abbildung 16: Aufbau der Ergebnisdatei
Die Datei enthält den Fitnesswert, der aus dem Verhältnis (ratio) und dem p-Wert (pValue) berechnet wird. Alle weiteren Angaben beziehen sich auf Proteine, die – gemäß den beiden Parametern – in der Messung ermittelt wurden, wobei sich alle Angaben nach dem „|“ ausschließlich auf die Proteine beziehen, die zusätzlich in der GeneCards vertreten sind. Weitere Angaben sind die normalisierte Pfadlänge (normPL); alle ermittelten Proteine (all); alle ermittelten Proteine, die nicht im Netzwerk vertreten sind (not@Net); und die Proteine in den einzelnen Ebenen (tier 0, tier 1, tier 2,…) des Netzwerkes bezogen auf das Startprotein bzw. die Startproteine (siehe 3.4).

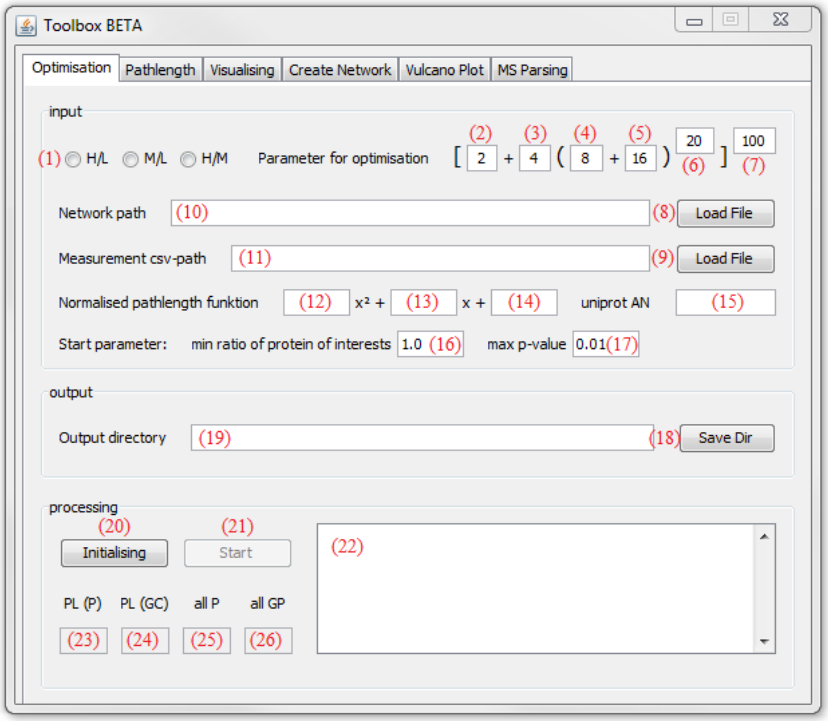


Abbildung 17: *Optimisation* Tab der Toolbox
Durch (1) kann zwischen den Ansätzen H/L, M/L und H/M gewählt werden. Die Parameter μ‘ (2), λ‘ (3), μ (4), λ (5), n (6) und m (7) können angepasst werden. Durch (8) kann die Datenbank und durch (9) kann die CSV-Datei geladen werden, wobei (10) und (11) die Pfade zu beiden Dateien darstellen, die auch manuell editiert werden können. In (12-14) sind die Terme der normalisierten Pfadlänge einzutragen und in (15) die UniProtKB-AN für das Protein, dessen Messung optimiert werden soll. Durch (16) und (17) können die Startparameter (Verhältnis und p-Wert) angepasst werden. Mit (18) kann der Pfad für die Ausgabedateien bestimmt werden. Dieser Pfad wird unter (19) dargestellt, der manuell editiert werden kann. Mit (20) wird die Optimierung initialisiert und mit (21) wird die Optimierung gestartet. In (22) wird der Status der Optimierung, in (23) wird die mittlere Pfadlänge aller Proteine der Messung, in (24) wird die mittlere Pfadlänge aller Proteine der Messung, die ebenfalls in der GeneCards auftreten, angezeigt. (25) zeit die Anzahl aller Proteine der Messung und (26) zeigt die Anzahl aller Proteine, die ebenfalls in der GeneCards stehen an.

5 Ergebnisse

5.1 Korrelation von MINT Score mit IntAct Score

Damit die Möglichkeit besteht, die Scores der Interaktionen zusätzlich in die Fitnessfunktion einbauen zu können, wurde versucht, den MINT-Score und den IntAct-Score miteinander zu korrelieren, sodass beide Scores ineinander umgerechnet werden können. Insgesamt sind 167.356 Interaktionen in beiden Datenbanken zwischen 53.430 Proteinen verzeichnet. 81.130 Interaktionen davon sind nur in der IntAct Datenbank verzeichnet und 63.920 Interaktionen sind nur in der MINT Datenbank verzeichnet. Von den übrigen 22.306 Interaktionen haben 8.149 Interaktionen sowohl in der MINT Datenbank, als auch in der IntAct Datenbank einen Score. In Abbildung 18**Fehler! Verweisquelle konnte nicht gefunden werden.** sind beide Scores gegeneinander aufgetragen, wobei die Pearson-Korrelation dieser Scores 0,376 beträgt.

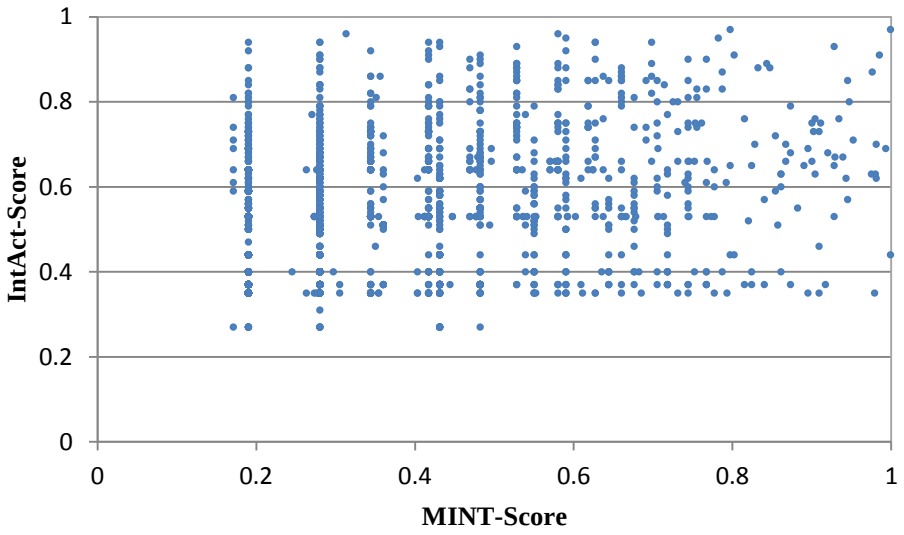


Abbildung 18: Korrelation zwischen MINT-Score und IntAct-Score
Es ist zu erkennen, dass zwischen den 8.149 Interaktionen, die sowohl von der IntAct, als auch von der MINT einen Score besitzen, keine signifikante Korrelation vorhanden ist.

5.2 Angaben zu den proteinzentrischen Netzwerken

Für diese Arbeit wurden drei Netzwerke – äquivalent zu den Proteinen, die untersucht wurden – erstellt: Das STAT-Netzwerk mit den Startproteinen STAT1 und STAT3, das BMI1-Netzwerk mit dem Startprotein BMI1 und das CDK9-Netzwerk mit dem Startprotein CDK9. Für die Proteine STAT1 und STAT3 wurde, wegen ihrer Fähigkeit miteinander ein Heterodimer auszubilden, ein einzelnes Netzwerk erstellt mit beiden Proteinen als Startproteine. Details zu den Netzwerken sind in Tabelle 6 aufgeführt. Für jedes der drei Netzwerke wurde, wie unter 3.5.1 beschrieben, die mittlere Pfadlänge zu zufällig aus dem Netzwerk entnommenen Proteinen bestimmt, woraus sich die Formeln zur Berechnung der normalisierten Pfadlängen ableiten lassen. Dies wurde ebenfalls für jedes Netzwerk separat bestimmt und ist auch in Tabelle 6 aufgeführt.

Tabelle 6: Angaben zu den erstellten Netzwerken
Alle drei Netzwerke enthalten die gleichen Proteine, weil alle 4 Startproteine über mehrere Proteine in den Datenbanken miteinander verbunden sind. Somit ist die Anzahl an Interaktionen aller drei Netzwerke ebenfalls gleich. Zusätzlich sind für jedes Netzwerk die mittlere Pfadlänge zu zufällig ausgewählten Proteinen sowie die daraus resultierende Formel der normalisierten Pfadlänge angeben.

	STAT-Netzwerk	BMI1-Netzwerk	CDK9-Netzwerk
Startprotein	STAT1, STAT3	BMI1	CDK9
Proteine 1. Ebene	53	16	20
Proteine 2. Ebene	748	375	144
Proteine 3. Ebene	4392	2675	2002
Proteine 4. Ebene	2920	4445	4561
Proteine 5. Ebene	470	992	1677
Proteine 6. Ebene	68	142	233
Proteine 7. Ebene	9	16	24
Proteine 8. Ebene	1	1	1
Proteine gesamt	8663	8663	8663
Interaktionen gesamt	Je 26002 (MINT: 2769 | MINT und IntAct: 3888 | IntAct: 19345)		
Mittlere Pfadlänge	3,3769	3,7524	3,9870
Formeln für die normierten Pfadlängen	$0{,}0293\ x^2$ $- 0{,}5492\ x + 1{,}5199$	$0{,}0183\ x^2$ $- 0{,}4501\ x + 1{,}4319$	$0{,}0122\ x^2$ $- 0{,}3958\ x + 1{,}3835$

5.3 Ergebnisse des Vorversuches

Der Vorversuch (siehe 3.2.1, Seite 22) diente als Grundlage der Fitnessfunktion. Dabei wurde ein Double-SILAC Ansatz genutzt, wobei eine *in situ* Biotinylierung und Affinitätsaufreinigungen von STAT3 durchgeführt wurde. Die Proteine CDK9 und STAT1 wurden beide quantifiziert, während BMI1 nicht quantifiziert werden konnte. In diesem Versuch wurden 1113 Proteine quantifiziert, die wenigstens in drei Replikaten gemessen wurden sowie keine Chaperone, keine Proteolysefaktoren und keine Proteine sind, die eine Domäne haben, die an Biotin binden kann. Von diesen 1113 Proteinen waren 221 in der GeneCards Datenbank unter STAT3 annotiert.

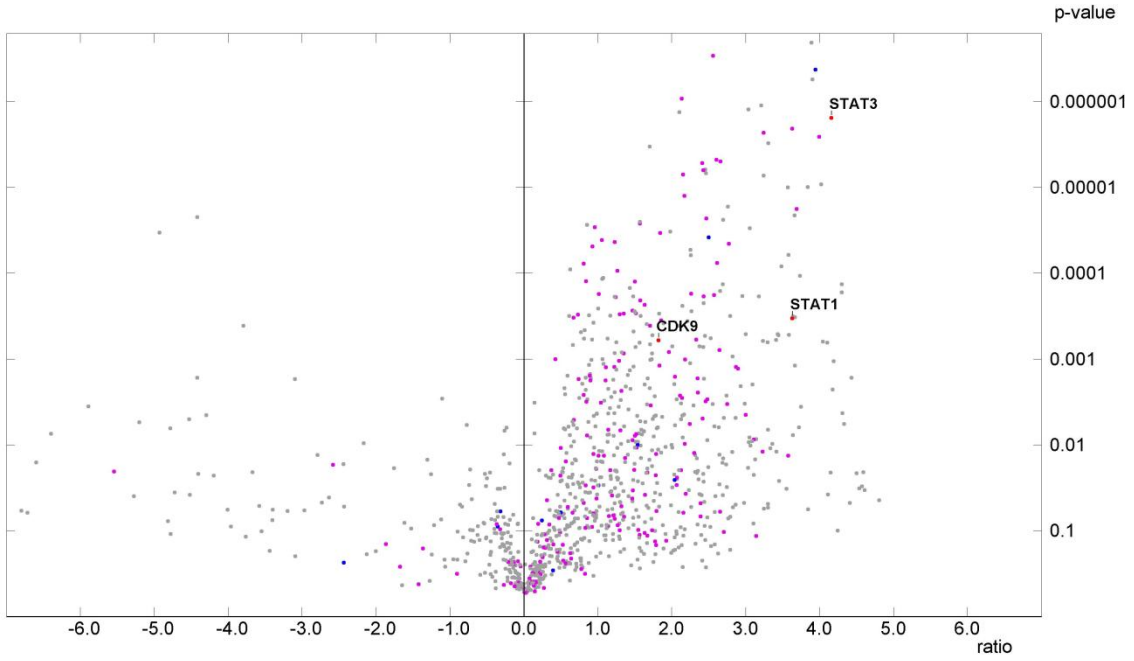


Abbildung 19: Vulcano Plot des Vorversuches
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die Proteine STAT3, STAT1 und CDK9 Rot markiert. Zusätzlich sind alle Proteine, die in der GeneCards Datenbank für STAT3 annotiert sind in Magenta und alle Proteine, die in der GeneCards Datenbank für STAT1 annotiert sind in Blau dargestellt.

5.4 Ergebnisse der Optimierungen

Bevor die Messungen optimiert werden konnten, musste, wie unter 4.6 beschrieben, eine Initialisierung erfolgen. Dabei wurde ermittelt, wie viele Proteine der Messung für die Optimierung genutzt werden konnten, wie viele von diesen wiederum in der GeneCards Datenbank auftreten und welche Werte die Pfadlängen sowie die normalisierten Pfadlängen von diesen beiden Proteinmengen haben. Für die Optimierung wurden nur Proteine genutzt, die wenigstens in zwei Replikaten gemessen wurden sowie keine Chaperone, keine Proteolysefaktoren oder Proteine, die eine Domäne haben, die an Biotin binden kann. Diese Angaben für alle acht Optimierungen sind in nachstehender Tabelle 7 aufgeführt.

Tabelle 7: Angaben zur Initialisierten der Optimierungen
Hierbei sind alle zur Initialisierung der Optimierung benötigten Angaben aufgeführt. Dabei handelt es sich um die Anzahl an verwendbaren Proteinen (P) sowie die Anzahl an verwendbaren Proteinen, die zusätzlich in der GeneCards Datenbank (GC) annotiert sind. Von diesen beiden Proteinmengen sind die Pfadlängen sowie die normierten Pfadlängen (PL) angeben.

	STAT3		STAT1		BMI1		CDK9	
Ansatz	H/L	M/L	H/L	M/L	H/L	M/L	H/L	M/L
Anzahl Proteine	817	809	1173	1222	1367	1232	1066	1048
Davon in der GC	221	223	35	34	9	7	345	346
Pfadlänge (P)	2,8905	2,8899	2,9444	2,9515	3,3648	2,0000	3,4875	3,4699
Pfadlänge (GC)	2,7488	2,7512	2,2941	2,2941	2,2500	3,3284	3,3304	3,3323
Normierte PL (P)	0,1772	0,1774	0,1568	0,1542	0,1246	0,1365	0,2007	0,1571
Normierte PL (GC)	0,2316	0,2307	0,4142	0,4142	0,5118	0,6049	0,2458	0,2000

Wird im Folgenden von bekannten Bindungspartnern gesprochen, handelt es sich um die Bindungspartner zu dem zu untersuchendem Protein, die aus den verwendeten Interaktionsdatenbanken entnommen wurden, mit denen das jeweilige PPIN erstellt wurde. Von potentiellen Bindungspartnern wird gesprochen, wenn zwei Proteine über ein Dreieck-Netzwerk-Motiv miteinander verbunden werden können und als vorhergesagte Bindungspartner werden alle Proteine bezeichnet, die von der PIP Datenbank als Bindungspartner zu dem jeweils zu untersuchendem Protein angegeben werden. Werden die gemessenen Proteine nach der Optimierung anhand entsprechender Werte für das Verhältnis und den p-Wert als Bindungspartner deklariert, wird von signifikanten Bindungspartnern gesprochen. Alle nachfolgenden Abbildungen, die einen Vulcano Plot darstellen, wurden mit dem unter 4.2 beschriebenen Softwaretool angefertigt.

5.4.1 Ergebnisse der Optimierung des STAT3 Experimentes

Die Optimierung des H/L Ansatzes der STAT3 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 1,9667 bei einem p-Wert von ≤ 0,333 aufweisen. In diesem Bereich wurden 16 Proteinen gemessen, von denen zwei in der GeneCards (STAT3 selbst und AHNAK) und eines in der PIPs (NUP153) annotiert sind. Auffällig ist, dass die drei bekannten Interaktionspartner zu STAT3, die in dieser Messung festgestellt werden konnten, nicht angereichert genug sind, um von dem Algorithmus als signifikant eingestuft werden zu können. Diese drei Proteine sind STAT1, RPL11 und RPS9. Weiterhin wurden zehn Proteine gemessen, die von der PIPs Datenbank als Bindungspartner zu STAT3 vorhersagt wurden. Davon wurde jedoch nur einer als signifikanter Bindungspartner (NUP153) bestimmt. Von allen gemessenen Proteinen konnten durch die Dreieck-Netzwerk-Motive drei Proteine als potentielle Bindungspartner (DNAJA1, HSPA1A und MCM3) eingestuft werden, wovon aber keines als signifikant klassifiziert werden konnte. CDK9 und BMI1 wurden in der Messung nicht quantifiziert. Das gesamte Ergebnis ist im nachstehendem Vulcano Plot (Abbildung 20) visualisiert.

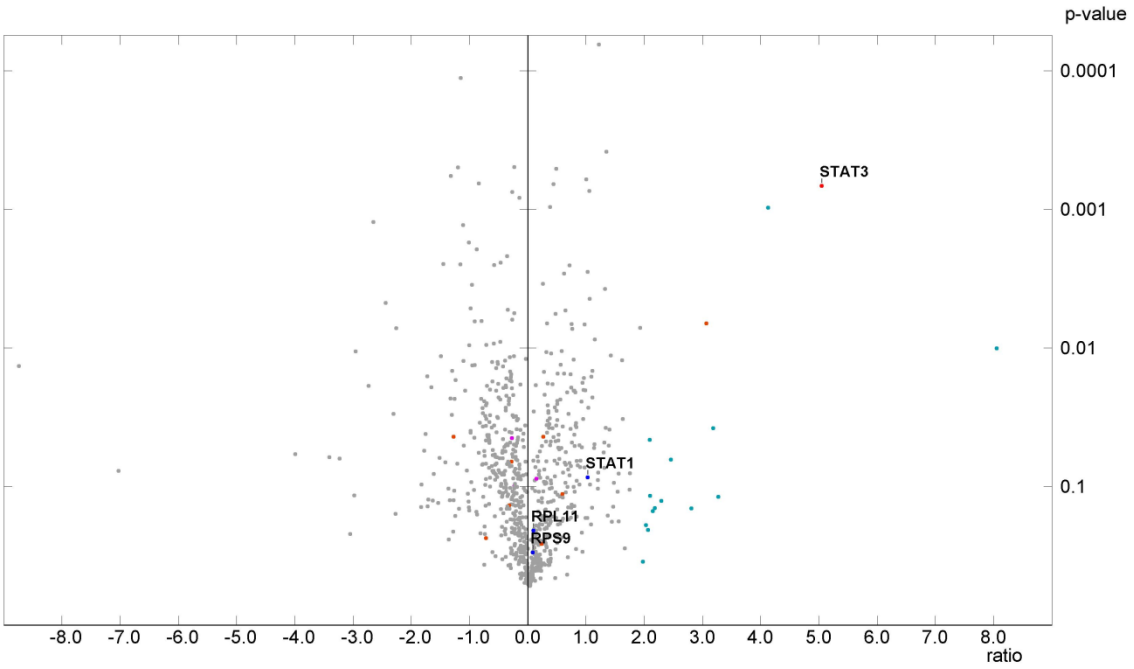


Abbildung 20: Vulcano Plot von STAT3, H/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von STAT3 Blau, die durch komplementäre Daten vermuteten Interaktionspartner Magenta und die von der PIPs Datenbank als Bindungspartner von STAT3 vorhergesagten Proteine (ausgenommen STAT1 als bekannter Bindungspartner und MCM3 als vorhergesagter Bindungspartner) sind Orange markiert. STAT3 selbst ist Rot hervorgehoben und die – gemäß den optimierten Parametern – als signifikant bestimmten Interaktionspartner Cyan dargestellt.

Die Optimierung des M/L Ansatzes der STAT3 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 0,9348 bei einem p-Wert von ≤ 0,083 aufweisen. In diesem Bereich wurden 14 Proteine quantifiziert, von denen zwei in der GeneCards (STAT3 selbst und AHNAK) und eines in der PIPs (NUP153) annotiert sind. Als signifikante Bindungspartner wurden in beiden Ansätzen die Proteine STAT3, AHNAK, NUP153, PDHA1 und INADL quantifiziert. Alle drei direkten Interaktionspartner zu STAT3 (STAT1, RPL11 und RPS9), die in diesem Versuch quantifiziert werden konnten, sind leicht abgereichert. Mittels Dreieck-Netzwerk-Motiven konnten zwei Proteine als potentielle Interaktionspartner bestimmt werden, wovon MCM3 nur schwach angereichert und HSPA1A deutlich abgereichert ist. Weiterhin wurden, von allen in der PIPs Datenbank als Bindungspartner zu STAT3 vorhergesagten Proteinen, sechs in diesem Versuch quantifiziert. Diese Proteine sind NUP153, MCM3, GNB2L1, STAT1, PHB2 und CSK. Dabei wurde MCM3 über die Dreieck-Netzwerk-Motive als potentieller Bindungspartner bestimmt. Von STAT1 ist bereits bekannt, dass es ein Bindungspartner zu STAT3 darstellt. Das gesamte Ergebnis ist im nachstehendem Vulcano Plot (Abbildung 21) visualisiert. In keinem der beiden STAT1 Versuche (H/L und M/L) wurde CDK9 oder BMI1 quantifiziert.

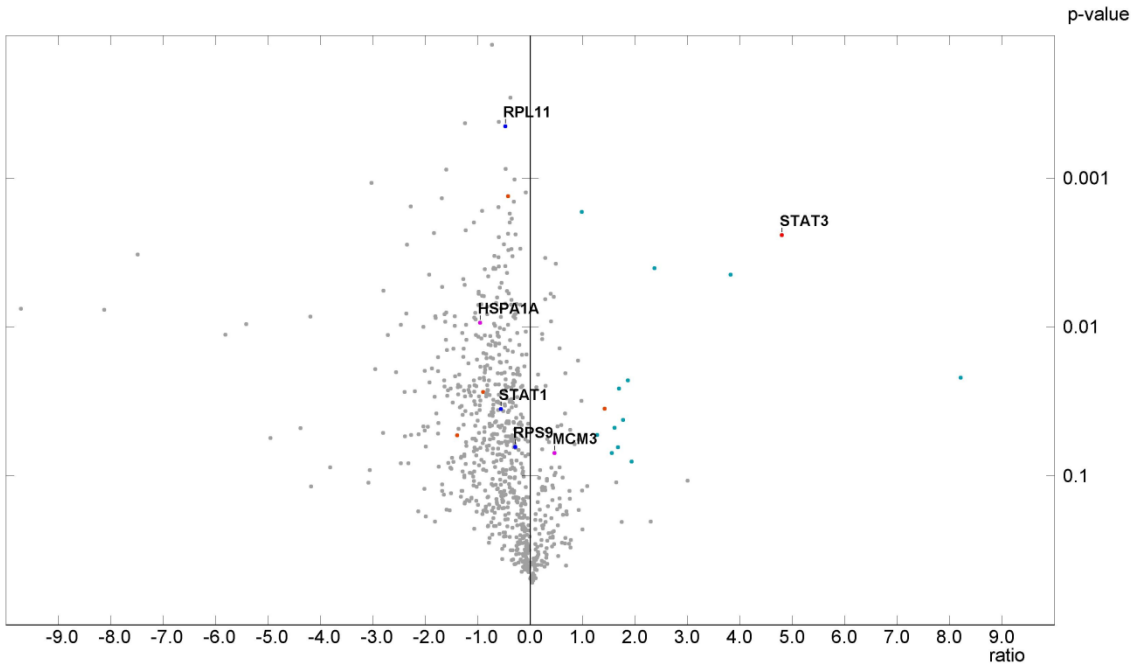


Abbildung 21: Vulcano Plot von STAT3, M/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von STAT3 Blau, die durch komplementäre Daten vermuteten Interaktionspartner Magenta und die von der PIPs Datenbank als Bindungspartner von STAT3 vorhergesagten Proteine (ausgenommen STAT1 als bekannter Bindungspartner und MCM3 als vorhergesagter Bindungspartner) sind Orange markiert. STAT3 selbst ist Rot hervorgehoben und die – gemäß den optimierten Parametern – als signifikant bestimmten Interaktionspartner Cyan dargestellt.

5.4.2 Ergebnisse der Optimierung des STAT1 Experimentes

Die Optimierung des H/L Ansatzes der STAT1 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 0,6605 bei einem p-Wert von ≤ 0,224 aufweisen. In diesem Bereich wurden 96 Proteinen gemessen, von denen Vier in der GeneCards Datenbank und der PIPs Datenbank (STAT3, STAT2, FGFR4 und TP53), Neun nur in der GeneCards Datenbank (STAT1, KPNA1, KPNA6, PLEC, SPTAN1, ACTN4, LDHB, GNB2L1 und SPTBN1) und Eines nur in der PIPs Datenbank (FGR) annotiert sind. Von den 96 Proteinen sind STAT3, STAT2, SPTAN1 und SPTBN1 als direkte Interaktionspartner von STAT1 bereits bekannt. Es wurden zwei weitere bekannte Bindungspartner zu STAT1 (DDB1 und GEMIN4) ebenfalls im Versuch gemessen, diese waren abgereichert. Ebenfalls abgereichert wurde CDK9 quantifiziert. In der gesamten Messung dieses STAT1 Ansatzes wurden insgesamt 15 Proteine quantifiziert, die von der PIPs Datenbank als Bindungspartner für STAT1 vorhergesagt wurden. Davon war Eines nicht signifikant genug und Neun waren abgereichert. Diese Ergebnisse sind im nachstehenden Vulcano Plot in Abbildung 22 visualisiert.

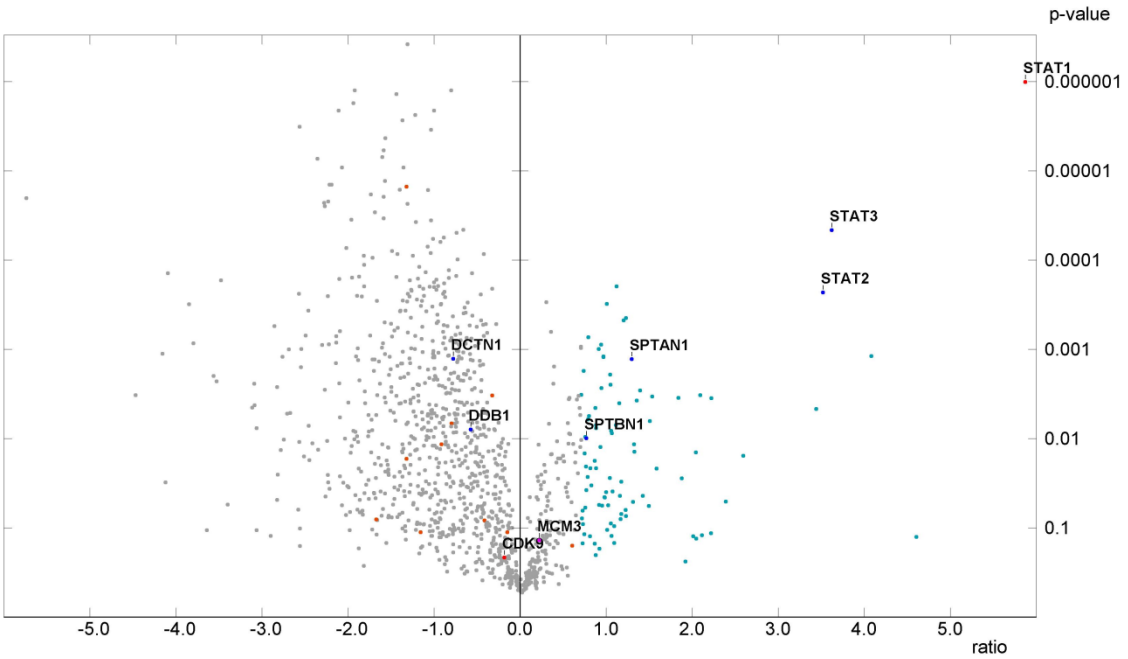


Abbildung 22: Vulcano Plot von STAT1, H/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von STAT1 Blau, der durch komplementäre Daten vermutete Interaktionspartner Magenta und die von der PIPs Datenbank als Bindungspartner von STAT1 vorhergesagten Proteine (ausgenommen STAT2 und STAT3 als bekannter Bindungspartner) sind Orange markiert. STAT1 selbst und CDK9 sind Rot hervorgehoben und die – gemäß den optimierten Parametern – als signifikant bestimmten Interaktionspartner Cyan dargestellt.

Die Optimierung des M/L Ansatzes der STAT1 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 1,0648 bei einem p-Wert von ≤ 0,1736 aufweisen. In diesem Bereich wurden 82 Proteinen quantifiziert, von denen eines in der GeneCards Datenbank und der PIPs Datenbank (STAT2), zwei nur in der GeneCards Datenbank (STAT1 selbst und GSTK1) und eines nur in der PIPs Datenbank (CSNK2A1) annotiert sind. Von den 82 Proteinen ist nur STAT2 als direkter Interaktionspartner von STAT1 bereits bekannt. Alle weiteren bekannten, quantifizierten Bindungspartner (SPTAN1, DDB1, DCTN1 und STAT3) waren nicht signifikant angereichert oder abgereichert. Von den insgesamt elf vorhergesagten Bindungspartnern von STAT1 durch die PIPs Datenbank wurden Neun abgereichert quantifiziert. Mit keinem quantifizierten Protein des M/L Ansatzes konnte ein Dreieck-Netzwerk-Motiv mit STAT1 gebildet werden. Diese Ergebnisse sind im nachstehenden Vulcano Plot in Abbildung 23 visualisiert.

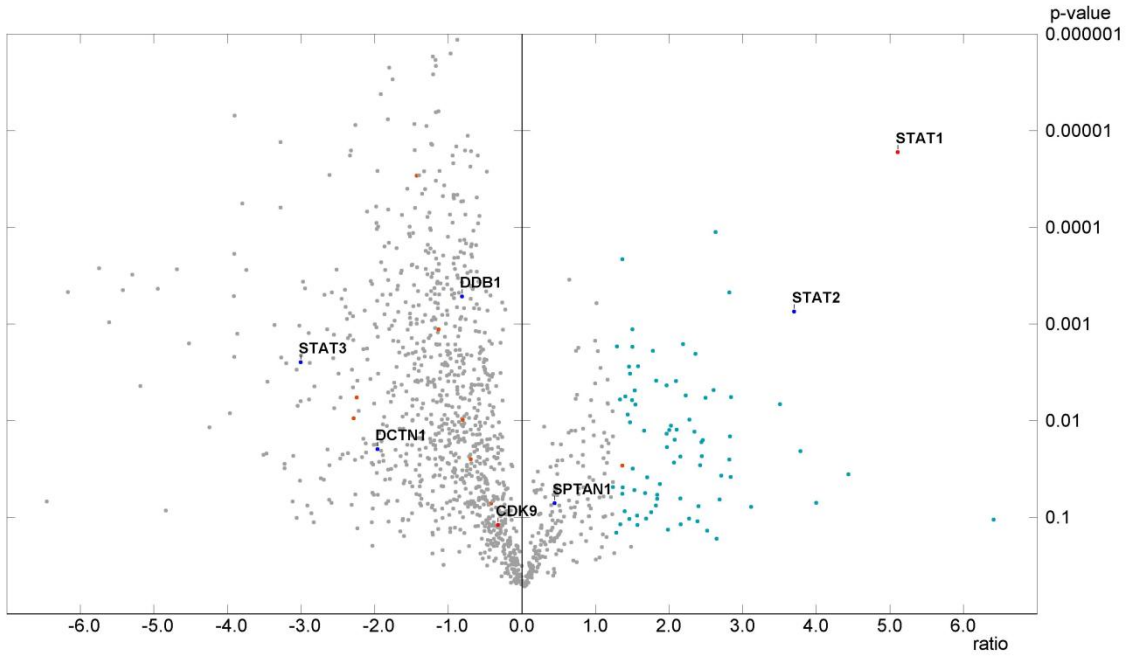


Abbildung 23: Vulcano Plot von STAT1, M/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von STAT1 Blau und die von der PIPs Datenbank als Bindungspartner von STAT1 vorhergesagten Proteine (ausgenommen STAT2 und STAT3 als bekannte Bindungspartner) sind Orange markiert. STAT1 selbst sowie CDK9 sind Rot hervorgehoben und die – gemäß den optimierten Parametern – in der Messung bestimmten Interaktionspartner Cyan dargestellt.

5.4.3 Ergebnisse der Optimierung des CDK9 Experimentes

Die Optimierung des H/L Ansatzes der CDK9 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 6,2881 bei einem p-Wert von ≤ 0,009 aufweisen. In diesem Bereich wurden die Proteine CDK9, CDC37, IL1RAPL1 und CCNT1 quantifiziert, wobei nur CCNT1 einen bereits bekannten Bindungspartner zu CDK9 darstellt und auch nur dieses Protein auf der GeneCards Datenbank als Bindungspartner zu CDK9 annotiert ist. Weitere bereits bekannte Bindungspartner, die in der Messung quantifiziert, aber als nicht signifikant eingestuft wurden, sind CASK, PPP1CC, LARP7 und HEXIM1. Das Protein AFF4 konnte als potentieller Bindungspartner zu CDK9 mittels Dreieck-Netzwerk-Motiven ermittelt werden, wobei auch dieses Protein als nicht signifikant von dem Optimierungsalgorithmus eingestuft wurde. Von den Proteinen CDK1 und CDK12 konnten alle drei Verbindungen durch komplementäre Daten zu CDK9 belegt werden. Jedoch sind zwischen CDK9 und CDK1 bzw. CDK9 und CDK12 keine Nachbarn zweiter Stufe im verwendeten PPIN vorhanden, sodass keine Dreieck-Netzwerk-Motive erstellt werden konnte. In diesem Versuch ist CDK1 leicht abgereichert und das Protein CDK12 ist relativ stark angereichert. Von CDK9 sind keine Daten auf der PIPs Datenbank vorhanden. Diese Ergebnisse dieses Versuches sind im nachstehenden Vulcano Plot in Abbildung 24 visualisiert.

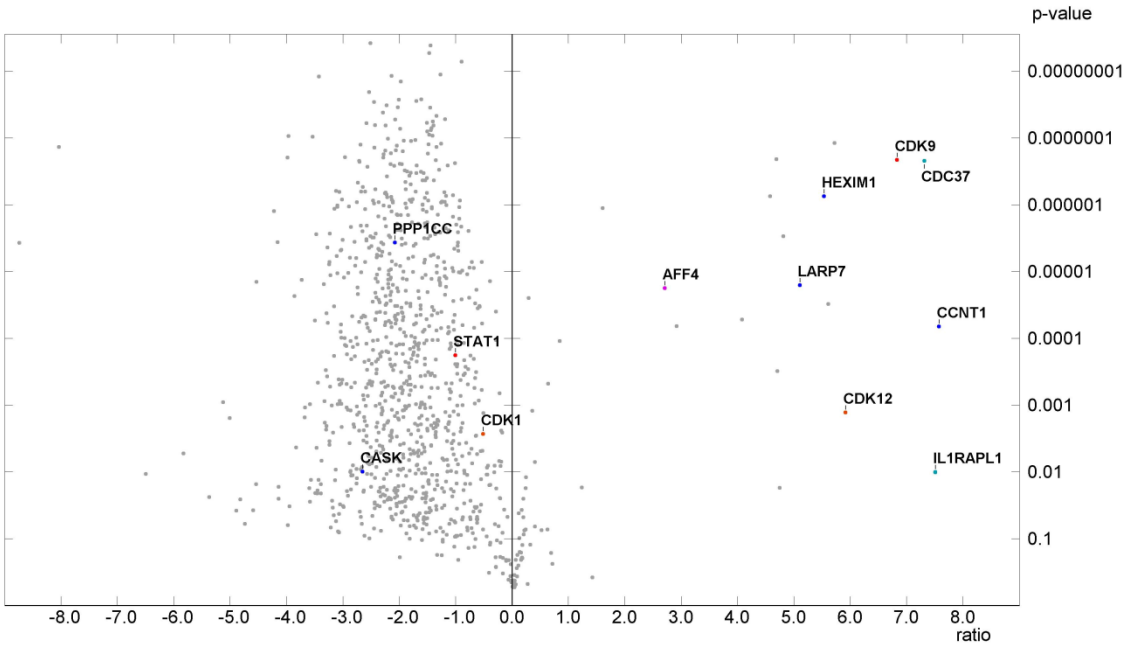


Abbildung 24: Vulcano Plot von CDK9, H/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von CDK9 Blau und die – gemäß den optimierten Parametern – in der Messung bestimmten Interaktionspartner Cyan dargestellt. CDK9 selbst und STAT1 sind Rot hervorgehoben. Die beiden Proteine CDK1 und CDK12 sind Orange dargestellt.

Die Optimierung des M/L Ansatzes der CDK9 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 5,2039 bei einem p-Wert von ≤ 0,0151 aufweisen. In diesem

Bereich wurden die Proteine CDK9, CDC37 und CCNT1 quantifiziert, wobei nur CCNT1 einen bereits bekannten Bindungspartner zu CDK9 darstellt und auch nur dieses Protein auf der GeneCards Datenbank als Bindungspartner zu CDK9 deklariert wurde. Weitere bereits bekannte Bindungspartner, die in der Messung quantifiziert, aber als nicht signifikant eingestuft wurden, sind PPP1CC, LARP7 und HEXIM1. Das Protein AFF4 konnte auch in diesem Versuch als potentieller Bindungspartner zu CDK9 mittels Dreieck-Netzwerk-Motiven ermittelt werden, wobei auch dieses Protein als nicht signifikant von dem Optimierungsalgorithmus eingestuft wurde. Die Proteine CDK1, welches leicht abgereichert quantifiziert wurde, und CDK12, welches stark angereichert quantifiziert wurde, wurden auch in diesem Ansatz quantifiziert. Diese Ergebnisse dieses Versuches sind im nachstehenden Vulcano Plot in Abbildung 25 visualisiert.

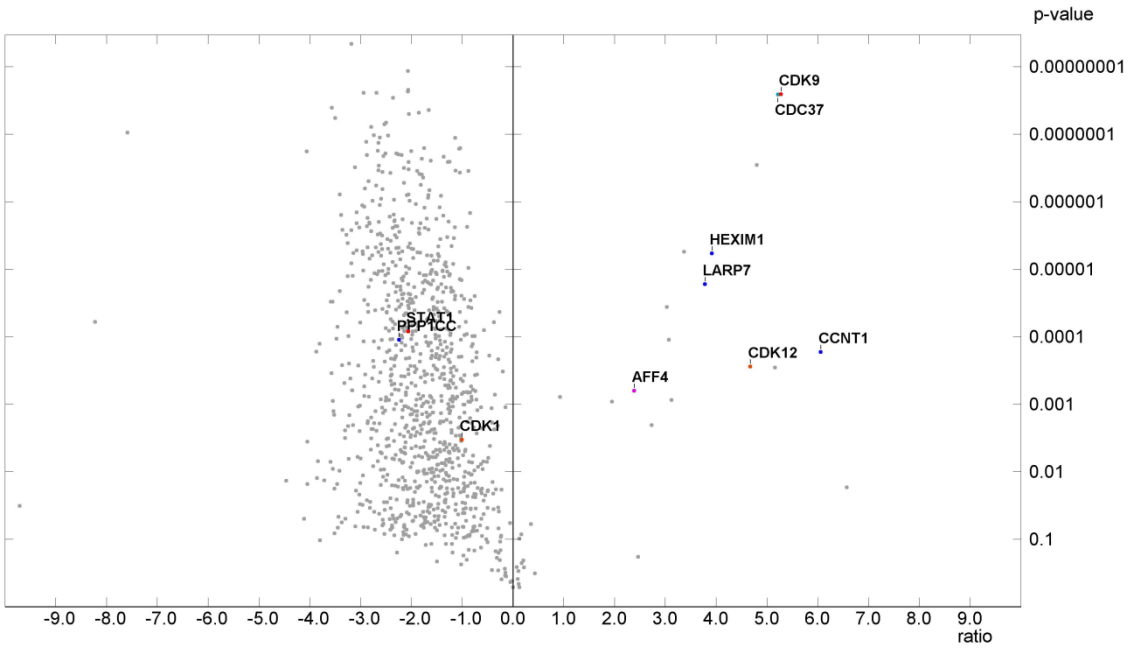


Abbildung 25: Vulcano Plot von CDK9, M/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von CDK9 Blau und die – gemäß den optimierten Parametern – in der Messung bestimmten Interaktionspartner Cyan dargestellt. CDK9 selbst und STAT1 sind Rot hervorgehoben. Die beiden Proteine CDK1 und CDK12 sind in Orange dargestellt.

5.4.4 Ergebnisse der Optimierung des BMI1 Experimentes

Die Optimierung des H/L Ansatzes der BMI1 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 1,5363 bei einem p-Wert von ≤ 0,0236 aufweisen. In diesem Bereich wurden 124 Proteine quantifiziert, von denen zwei in der GeneCards Datenbank (RING1 und RNF2) annotiert und gleichzeitig bekannte Bindungspartner zu BMI sind. Die PIPs Datenbank enthält keine Informationen zu BMI1 und mittels Dreieck-Netzwerk-Motiven können ebenfalls keine zusätzlichen potentiellen Bindungspartner ermittelt werden. Die Proteine CDK9 und STAT1 sind hingegen stark abgereichert. STAT3 wurde in diesem Versuch nicht quantifiziert. Diese Ergebnisse sind im nachstehenden Vulcano Plot in Abbildung 26 visualisiert.

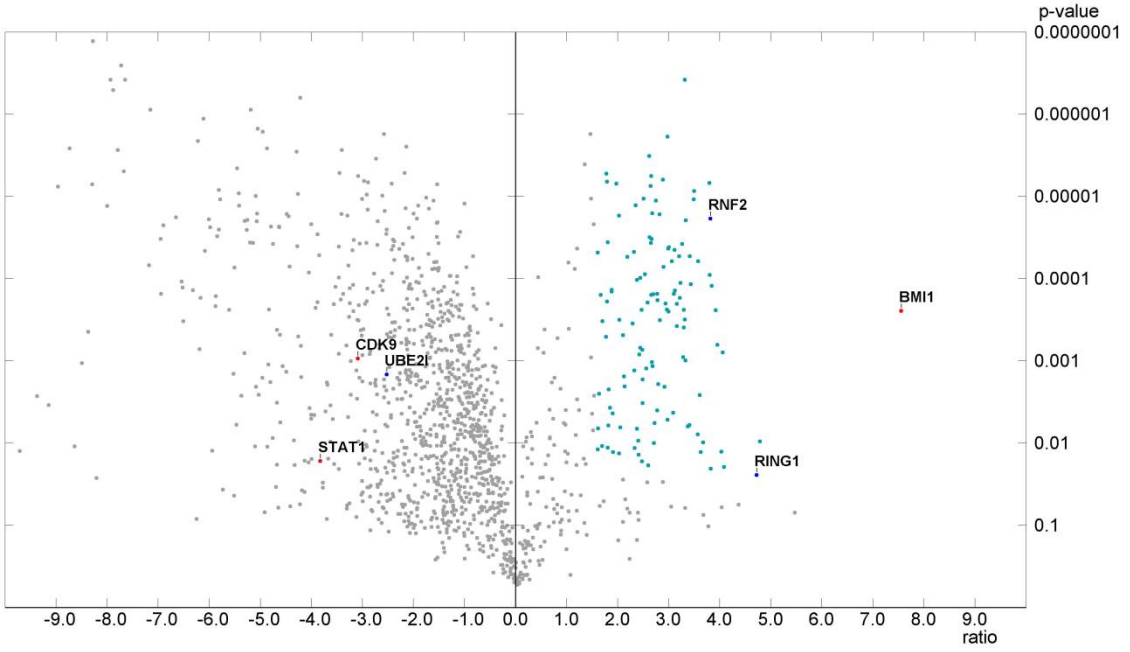


Abbildung 26: Vulcano Plot von BMI1, H/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von BMI1 Blau und die – gemäß den optimierten Parametern – in der Messung bestimmten Interaktionspartner Cyan dargestellt. BMI1 selbst sowie CDK9 und STAT1 sind Rot hervorgehoben.

Die Optimierung des M/L Ansatzes der BMI1 Experimente ermittelte, dass die signifikanten Bindungspartner ein Verhältnis von ≥ 0,0724 bei einem p-Wert von ≤ 0,4184 aufweisen. In diesem Bereich wurden 46 Proteine quantifiziert, wobei erneut die beiden Proteine RING1 und RNF2 (bekannte Bindungspartner zu BMI1 und in der GeneCards Datenbank zu BMI1 annotiert) vertreten sind. In diesem Ansatz konnten ebenfalls keine Dreieck-Netzwerk-Motive ermittelt werden. STAT1 ist hingegen stark abgereichert. STAT3 und CDK9 wurden in diesem Versuch nicht quantifiziert. Diese Ergebnisse sind im Vulcano Plot in Abbildung 27 visualisiert.

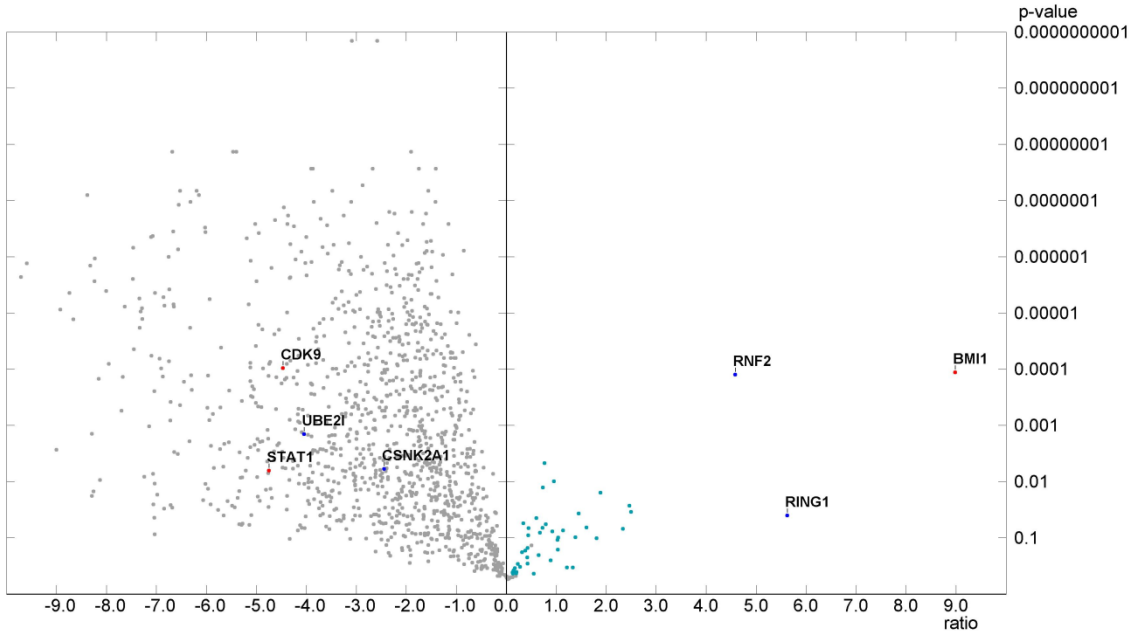


Abbildung 27: Vulcano Plot von BMI1, M/L
In dieser Abbildung ist das Verhältnis auf der Abszisse und der p-Wert auf der Ordinate aufgetragen. Hierbei sind die bereits bekannten Interaktionspartner von BMI1 Blau und die – gemäß den optimierten Parametern – in der Messung bestimmten Interaktionspartner Cyan dargestellt. BMI1 selbst sowie STAT1 sind Rot hervorgehoben.

6 Diskussion

6.1 Fitnessfunktion

Die Fitnessfunktion orientiert sich anhand der Daten des verwendeten PPINs, das zur Optimierung der jeweiligen Messung verwendet wird. Sie bezieht das Netzwerk derart ein, dass Interaktionspartner bevorzugt werden, deren Pfadlänge zu dem zu untersuchenden Protein bzw. zu den zu untersuchenden Proteinen (das Zentrum des jeweiligen PPINs) besonders kurz ist. Diese Aufgabe nimmt der Pfadlängenterm ein. Weiterhin wurde geprüft, ob die quantifizierten Proteine bisher mit dem zu untersuchendem Protein bzw. mit den zu untersuchenden Proteinen in Relation gebracht werden konnten. Dafür wurde die Metadatenbank GeneCards verwendet, die aus unterschiedlichen Datenquellen bekannte, potentielle und vorhergesagte Bindungspartner zu einem Protein angibt. Dieser Aufgabe ist der Genauigkeitsterm der Fitnessfunktion zugeteilt, der die Proteine aus der GeneCards Datenbank bevorzugt, die eine kurze Pfadlänge zu dem zu untersuchendem Protein haben. Der Vollständigkeitsterm sollte abschließend verhindern, dass nur einige wenige, bereits bekannte Proteine von der Optimierung als signifikant eingestuft werden. Als Grundlage der gesamten Fitnessfunktion und im Besonderen als Grundlage des Vollständigkeitsterms dienten die Daten des unter 3.2.1 beschriebenen Vorversuches. In diesem wurden ca. 15 % der quantifizierten Proteine als potentielle Interaktionspartner eingestuft.

Schwächen der Fitnessfunktion liegen darin, dass sie anhand der Daten des Vorversuchs angefertigt und diese Daten als Testdatensatz verwendet wurden. Die Daten dieses Vorversuchs wurden allerdings unter Verwendung von Doppel-SILAC gewonnen. Die in dieser Arbeit durch die Fitnessfunktion optimierten Daten wurden dagegen unter Verwendung von Triple-SILAC gewonnen. Ausgewertete Triple-SILAC Daten, die als Testdatensatz verwendet werden konnten, waren nicht vorhanden. Weiterhin besitzen die von der GeneCards Datenbank aufgelisteten Proteine keinen Score, der deren Relevanz verdeutlicht. Somit ist von den in der GeneCards Datenbank annotierten Proteinen nicht erkennbar, ob es sich um richtige Interaktionspartner oder vorhergesagte Interaktionspartner handelt. Folglich wurden Beide gleichermaßen gewichtet.

Ein Verbesserungsansatz für die Fitnessfunktion war das Einbinden der von den Interaktionsdatenbanken vergebenen Scores zu den einzelnen darin aufgeführten Interaktionen. Beide Scores müssten dafür korreliert werden, damit eine Vergleichbarkeit hergestellt werden konnte. Aus diesem Grund wurden nur zwei Interaktionsdatenbanken zur Erstellung des PPINs verwendet, die IntAct und die MINT. Der Versuch, die von beiden Datenbanken für die einzelnen PPIs vergebenen Scores in die Berechnung einfließen zu lassen, scheiterte aber an der nicht signifikanten Korre-

lation, die zwischen beiden Scores ermittelt werden konnte. Die Ursache für diese Unterschiede der Scores, die grundlegend einen ähnlichen Aufbau haben, liegt daher wahrscheinlich weniger in den Berechnungen der Werte selbst, als in den Daten, die den Datenbanken im Einzelnen von einer Interaktion übermittelt werden. Sind die Daten von der gleichen Interaktion unterschiedlich, die beiden Datenbanken übermittelt werden, entstehen selbst bei gleicher Berechnung der Scores unterschiedliche Werte. Weitere Verbesserungsansätze der Fitnessfunktion sind im Ausblick (Gliederungspunkt 7) gegeben.

6.1.1 Verwendete Protein-Protein Interaktionsnetzwerke

Anhand der Daten aus Tabelle 6 (siehe 5.2, Seite 46) geht hervor, dass alle drei verwendeten PPINs (das STAT-PPIN, das BMI1-PPIN und das CDK9-PPIN) grundlegend die gleichen Netzwerke sind, aber auf einem anderen Startprotein bzw. auf anderen Startproteinen aufbauen. Die drei Netzwerke enthalten 26002 Interaktionen zwischen 8663 Proteinen, wobei die IntAct Datenbank die größere Menge an Interaktionen zu dem Netzwerk beisteuert. Aus den ersten Ebenen der Netzwerke geht hervor, dass das STAT-PPIN die höchste Dichte an Interaktionen um beide Startproteine aufweist. Die Angaben zu den mittleren Pfadlängen, die bei dem STAT-PPIN am kürzesten ist, unterstreichen diese Gegebenheit. Gründe dafür sind, dass dieses PPIN im Gegensatz zu den anderen beiden Netzwerken zwei Startproteine hat und das STAT Interaktom im Vergleich zu den anderen beiden Interaktomen bisher am besten erforscht scheint. Der Grund zur letzeren Annahme ist, dass die PIPs Datenbank keine bzw. wenige Angaben zu den Proteinen BMI1 und CDK9 besitzen, jedoch über eine Vielzahl an Informationen bzgl. STAT1 und STAT3 verfügt. Grundsätzlich kann davon ausgegangen werden, dass Proteine, die in dichten Bereichen eines Netzwerkes vorkommen, leichter mit Vorhersagemethoden untersucht werden können. Somit ist zu erwarten, dass die Analysen der beiden STAT Proteine eindeutigere Ergebnisse liefern, als die Analysen der Proteine BMI1 und CDK9.

6.2 Evaluierung der Optimierungsergebnisse

Bereits aus den Initialisierungsergebnissen lassen sich erste, wichtige Informationen zu den Versuchen gewinnen. So geht aus Tabelle 7 (siehe 5.4, Seite 48) hervor, dass bei den Messungen beider STAT3 Versuche (H/L und M/L) wesentlich weniger Proteine quantifiziert wurden, als in den Messungen der anderen drei Proteine. Das deutet darauf hin, dass die Probenqualität etwas schlechter als bei den anderen Versuchen gewesen sein könnte oder bei der Messung Ungenauigkeiten aufgetreten sein könnten. Wahrscheinlicher ist aber, dass bei STAT3 deshalb weniger Prote-

ine quantifiziert wurden, weil kein technisches Replikat von den Proben gemessen wurde. Somit konnten nur drei statt sechs Replikate in die Analyse einbezogen werden. Das hat zur Folge, dass ein Protein, das nur in einem der Triplikate auftaucht, auch nur einmal gemessen wird. Bei einem technischen Replikat wird dieses Protein zweimal gemessen und somit auch in die Analyse einbezogen.

Weiterhin ist zu erwarten, dass die mittleren Pfadlängen der Netzwerke größer sind, als die mittleren Pfadlängen aller in den jeweiligen Messungen quantifizierten Proteine. Grund dafür ist, dass mehr direkte Bindungspartner quantifiziert werden, als Proteine, die im Netzwerk sehr weit entfernt vom zu untersuchendem Protein bzw. von den zu untersuchenden Proteinen auftauchen. Diese Annahme wird auch durch alle Messungen bestätigt (Vergleich Tabelle 7 mit den mittleren Pfadlängen der in den Messungen quantifizierten Proteinen, Seite 48 und Tabelle 6 mit den mittleren Pfadlängen aller im Netzwerk enthaltenen Proteinen, Seite 46). Zusätzlich wird diese Annahme von den mittleren Pfadlängen, berechnet aus den quantifizierten Proteinen, die zusätzlich in der GeneCards Datenbank annotiert sind, bekräftigt. Hintergrund ist, dass in der GeneCards Datenbank Proteine annotiert sind, die zu den zu untersuchendem Protein bekannte, potentielle oder vorhergesagte Bindungspartner sind und daher im PPIN im Mittel eine kürzere Pfadlänge aufweisen sollten, als zufällig ausgewählte Proteine. Nur die Messung des M/L Ansatzes von BMI1 wird dieser Annahme nicht gerecht, wobei das durch die geringe Anzahl an gefundenen GeneCards Einträgen zu begründen ist. Sollte eine Messung eine größere mittlere Pfadlänge, als das zugrunde liegende Netzwerk aufweisen, ist anzunehmen, dass das zu untersuchende Proteine hinreichend gut untersucht ist oder bei der Probenanfertigung und Probenmessung erhebliche Fehler aufgetreten sind.

6.2.1 STAT3 Experiment

Bei dem Ergebnis fällt auf, dass alle drei quantifizierten Proteine, die bereits als Bindungspartner zu STAT3 bekannt sind (STAT1, RPL11 und RPS9) nach der Optimierung des H/L Ansatzes nicht zu den signifikanten Bindungspartnern zugeordnet wurden. Dies gilt ebenfalls für die durch Dreieck-Netzwerk-Motive bestimmten Proteine DNAJA1, HSPA1A und MCM3, die ein von den beiden bekannten Bindungspartnern RPL11 und RPS9 unwesentlich abweichendes Verhältnis haben. Der Algorithmus könnte aus diesem Grund, dass sowohl die drei bekannten, als auch die drei potentiellen Bindungspartner nicht zu den signifikanten Bindungspartnern gehören, das Ergebnis zu stringent gewählt haben. Eine Lösung für diesen Sachverhalt stellt das Nachbereiten des Ergebnisses durch die komplementären Daten dar bzw. das direkte Einbinden der komplementären Daten in den Algorithmus. STAT1 ist der Einzige bekannte Bindungspartner, der quantifiziert wurde und zu STAT3 eine Verbindung über die komplementären Daten besitzt. Somit lässt sich eine weniger stringente Auswahl der signifikanten Bindungspartner bis einschließlich STAT1 begründen. Insge-

samt sind die An- bzw. Abreicherungen der fünf Proteine (DNAJA1, HSPA1A und MCM3 als potentielle Bindungspartner und RPL11 und RPS9 als bekannte Bindungspartner) im Vergleich zu den im Vorversuch erkennbaren An- bzw. Abreicherungen sehr marginal. Auch in dem M/L Ansatz sind die An- bzw. Abreicherungen der benannten bzw. potentiellen Bindungspartner marginal, wobei DNAJA1 nicht quantifiziert wurde. Bemerkenswert ist, dass die drei quantifizierten bekannten Bindungspartner alle abgereichert sind. Die von der PIPs Datenbank vorhergesagten Bindungspartner besitzen ebenfalls keine klare Tendenz hinsichtlich ihrer Bindungsaffinität zu STAT3.

Grundsätzlich ist hierbei deutlicher ein Unterschied zwischen den Double-SILAC und Triple-SILAC anhand der an- und abgereicherten Proteine zu erkennen. Ursache dafür ist, dass nicht nur zwei Zelllysate vermischt werden, sondern drei. Die absolute Proteinmenge an Bindungspartnern, die in einem Ansatz aufgereinigt werden bleibt somit gleich, wobei die Menge an Proteinen im Zelllysat, die unspezifisch binden, zu nimmt und mehr Hintergrundrauschen entsteht.

6.2.2 STAT1 Experiment

Im Vergleich zu den STAT3 Versuchen sind die STAT1 Experimente weniger marginal ausgefallen. Dies belegen die Ergebnisse, die STAT3 in beiden Ansätzen des STAT1 Versuches erzielt hat. So ist es im H/L Ansatz signifikant angereichert und im M/L Ansatz stark abgereichert, wie es anhand Literaturreferenzen erwartet wurde. Weiterhin ist STAT2, ein wichtiger Bindungspartner zu STAT1, in beiden Ansätzen als signifikanter Bindungspartner Identifiziert worden. Die Interaktion zwischen STAT1 und STAT2 ist in beiden verwendeten Datenbanken (IntAct und MINT) vorhanden und zusätzlich ist zwischen beiden eine Literatur Kookkurrenz und Domänen-Domänen Interaktionen vorhanden. Zusätzlich ist erwähnenswert, dass im H/L Ansatz alle vier quantifizierten, bekannten Interaktionspartner, die angereichert sind, als signifikant eingestuft wurden. Mit MCM3 wurde im H/L Ansatz ein potentieller Bindungspartner durch die Dreieck-Netzwerk-Motive ermittelt, der als nicht signifikant eingestuft werden konnte, weil es nur marginal angereichert quantifiziert wurde. CDK9 wurde in beiden Ansätzen abgereichert quantifiziert, obwohl es mit dem STAT-Komplex funktionell in Verbindung steht. Sowohl im H/L Ansatz, als auch im M/L Ansatz ist zu erkennen, dass die vorhergesagten Bindungspartner durch die PIPs Datenbank nahezu alle abgereichert quantifiziert wurden. Die PIPs Datenbank stellt demnach auch für die STAT1 Experimente keine verlässliche Evaluierungsgrundlage dar. SPTAN1 ist der einzige bekannte Bindungspartner, der im M/L Ansatz als nicht signifikant eingestuft wurde. Er besitzt keine von komplementären Daten implizierte Verbindung zu STAT1, weshalb diese Klassifizierung gerechtfertigt ist. Zusammenfassend lässt sich sagen, dass die Optimierung der STAT1 Experimente keiner Nachbesserung bedarf.

6.2.3 CDK9 Experiment

Das Experiment von CDK9 ergab eine geringe Anzahl an angereicherten Proteinen, wodurch in diesem Fall eine automatische Auswertung nahezu übermotiviert erscheint. Leider ergab die Optimierung eine extrem stringente Lösung vor mit nur drei signifikanten Bindungspartnern im H/L Ansatz und zwei signifikanten Bindungspartnern im M/L Ansatz. Dieser Fall stellt gleichzeitig jedoch ein sehr gutes Beispiel dar, wie komplementäre Daten und Dreieck-Netzwerk-Motive die Ergebnisse der Optimierung verbessern können. AAF4 ist ein Protein, das ein derartiges Motiv mit CDK9 bildet, und in beiden Ansätzen (H/L und M/L) quantifiziert wurde. Durch das Dreieck-Netzwerk-Motiv kann davon ausgegangen werden, dass ein Verhältnis auf dem Level von AAF4 in beiden Ansätzen ausreicht, um signifikante Bindungspartner zu klassifizieren. Zusätzlich können zwischen CDK9 und CDK1 bzw. CDK9 und CDK12 jeweils für alle drei komplementären Datentypen eine Verbindung ermittelt werden. Jedoch wurden in dem verwendetem Netzwerk zwischen CDK9 und CDK1 bzw. CDK9 und CDK12 keine Nachbarn zweiter Stufe ermittelt, sodass auch kein Dreieck-Netzwerk-Motiv gebildet werden konnte. CDK12 ist in beiden Ansätzen stärker angereichert als AAF4. Durch die Verbindung, die alle drei komplementären Datentypen zwischen CDK9 und CDK12 bilden, lässt es sich zusätzlich nicht ausschließen, dass CDK12 kein signifikanter Bindungspartner sein soll. Selbst für CDK1, das in diesem Experiment abgereichert gefunden wurde, ist eine Verbindung über drei komplementäre Datentypen zu deutlich, als das man es ganz als Bindungspartner zu CDK9 ausschließen könne. Grundsätzlich deuten aber beide Fälle darauf hin, dass das verwendete Netzwerk nicht vollständig ist. Wie eingangs erwähnt, ist die Unvollständigkeit von PPINs ein bekanntes Problem bei der Analyse von Interaktomen. Dem kann etwas entgegen gewirkt werden, indem mehr Interaktionsdatenbanken zur Erstellung des Netzwerkes genutzt werden, damit zwischen CDK9 und CDK1 bzw. CDK9 und CDK12 eine Interaktion mittels Dreieck-Netzwerk-Motiven vorhergesagt werden kann.

6.2.4 BMI1 Experiment

Es konnten zu diesem Protein weder Dreieck-Netzwerk-Motive gebildet werden, noch ist es in der PIPs Datenbank vertreten. Somit ist die Evaluierung der Optimierung kaum möglich. Von dem Protein wurden vier Proteine (RNF2, RING1, UBE2I und CSNK2A1) im H/L Ansatz bzw. drei Proteine (RNF2, RING1 und UBE2I) im M/L Ansatz, die bekannte Interaktionspartner zu BMI1 sind, quantifiziert. Dabei ist auffällig, dass RNF2 und RING1 in der GeneCards zu BMI1 annotiert sind und das BMI1 mit beiden Proteinen Literatur Kookkurrenzen sowie GO-Term Ähnlichkeiten aufweist. Zusätzlich sind in den jeweiligen Datenbanken für beide Proteine relativ hohe Scores angegeben (0,59 für RNF2, MINT; 0,72 für RING1, IntAct). UBE2I besitzt mit 0,37 (IntAct) dagegen nur einen Score, der knapp über dem gewählten Schwellwert liegt. Genau dieses UBE2I ist in

beiden Ansätzen stark abgereichert quantifiziert worden. Somit geben die komplementären Daten in diesem Fall einen Hinweis, dass es sich bei diesem Protein um einen falsch positiven Interaktionspartner handeln könnte. Weiterhin ist eine manuelle Korrektur des M/L Ansatzes erforderlich, aufgrund der Klassifizierung von selbst relativ marginal angereicherten Proteinen als signifikante Bindungspartner. Dabei würde die Bestimmung von signifikanten Bindungspartnern, aufgrund von fehlenden komplementären Informationen, lediglich auf die beiden bekannten Bindungspartner RING1 und RNF2 sowie BMI1 selbst fallen. Dies ist vor Allem darin begründet, dass diese drei Proteine wesentlich stärker angereichert sind, als alle anderen quantifizierten Proteine.

Das Protein BMI wurde für die Analyse des STAT1/STAT3 Interaktoms als negative Vergleichsprobe ausgewählt. In beiden Ansätzen (H/L und M/L) konnte sowohl für STAT1, als auch für CDK9 eine starke Abreicherung quantifiziert werden, sodass die Verwendung von CDK9 als Vergleichsprobe gerechtfertigt ist. Werden dennoch Interaktionspartner von BMI1 und STAT1, STAT3 oder CDK9 ermittelt, könnten Diese falsch positive Interaktionspartner darstellen.

6.3 Ergebnisse der automatischen Prozessierung

So ist festzustellen, dass die Optimierung für beide optimieren Ansätze von STAT1, den M/L Ansatz von STAT3 sowie dem H/L Ansatz von BMI1 Ergebnisse geliefert hat, die auch durch das nachträgliche, manuelle Integrieren der komplementären Daten nicht verbessert werden konnten. Somit ist anzunehmen, dass die Optimierung für diese Ansätze funktioniert hat. Nach Einbeziehung der komplementären Daten ist bei beiden Ansätzen von CDK9 und dem H/L Ansatz von STAT3 aufgefallen, dass sie eine scheinbar zu stringente Auswahl getroffen haben. Dies ist speziell bei den CDK9 Ansätzen aufgefallen. Durch Visualisierung der Vulcano Plots zu den einzelnen Ansätzen und dem Hinzufügen der Dreieck-Netzwerk-Motive kann dieser Umstand jedoch deutlich gemacht und korrigiert werden. Weiterhin wurde für den M/L Ansatz von BMI1 eine relativ unstringente Auswahl getroffen, die selbst marginalste Anreicherungen als signifikant klassifiziert. Ursache für Ungenauigkeiten in der Auswahl an signifikanten Bindungspartnern ist ein Mangel an Informationen. Das fällt besonders bei der Optimierung des M/L Ansatz von BMI1 auf, der einzige Ansatz, bei dem sehr marginal angereicherte Proteine als signifikant klassifiziert wurden. Gleichzeitig stellt dieser Umstand aber auch eine mögliche Lösung dar. Dreieck-Netzwerk-Motive und Komplementäre Daten können ebenfalls bereits von den aus den Messungen erhaltenen Daten erstellt und prozessiert werden. Einzig eine Integrierung in die Fitnessfunktion ist noch nicht erfolgt. Dabei ist anzuraten, nicht nur die Dreieck-Netzwerk-Motive selbst, sondern auch direkt die komplementären Daten zu integrieren. Dadurch ist es beispielsweise auch möglich einen dynamischen und genaueren Score aus den GO-Termen anstelle der Schwellwerte zu implementieren.

6.4 Vergleich mit komplementären Ansätzen

Zusammenfassend ist weiterhin fest zu halten, dass die PIPs Datenbank für eine Evaluierung der Bindungspartner zu wenige Daten besitzt. Lediglich für STAT3 und STAT1 waren Daten vorhanden. Diese vorhergesagten Proteine wurden aber in den seltensten Fällen angereichert quantifiziert bzw. als signifikant eingestuft. Die Schwäche der PIPs Datenbank im Gegensatz zu den Dreieck-Netzwerk-Motiven ist, dass sie auch einen Score erstellt, wenn keine Topologiedaten aus einem partiell erzeugten Netzwerk um das zu untersuchende Protein vorhanden sind. Somit stützen sich einige vorhergesagte Informationen nur auf komplementäre Daten.

Weiterhin sind in Anhang B die GeneCards Annotierungen zu jedem der vier experimentell untersuchten Proteine aufgetragen. Dabei ist keine signifikante Verteilung zu erkennen. Dies geht ebenfalls aus den Ergebnissen des Vorversuches (Abbildung 19) hervor. Dieser Umstand ist mit dem Meta-Datenbank Gedanken an sich zu erklären. Diese Datenbank enthält nicht nur experimentell untersuchte Interaktionen, sondern auch vorhergesagte und assoziierte. Bei der Optimierung werden jedoch alle in der GeneCards Datenbank, die unter einem zu untersuchenden Protein annotiert sind, mit in die Berechnung einbezogen. Somit ist auf lange Sicht eine Abkehr von der GeneCards Datenbank für die Optimierungsengine anzuraten.

7 Zusammenfassung und Ausblick

Zusammenfassend kann aus den Ergebnissen der vorliegenden Forschungsarbeit geschlossen werden, dass eine Optimierung von Proteininteraktionsdaten, gewonnen durch einer Strategie aus Triple-SILAC, verbunden mit *in situ* Biotinylierung und Affinitätsmassenspektrometrie möglich ist. Zu diesem Zweck wurde eine Software entwickelt, die eine schnelle Bearbeitung, Verarbeitung und auch Visualisierung der Daten ermöglicht. Die gewonnenen Ergebnisse durch die Optimierungsengine bedürfen dennoch einer Nachevaluierung durch komplementäre Daten, um genauere Ergebnisse zu erzielen.

Dieser Ansatz versucht fehlende Informationen von Interaktomen zu kompensieren, indem es bereits vorhandene Proteininteraktionsdaten aus verschiedenen Datenbanken mit komplementären Informationen, basierend auf Daten aus Meta-Datenbanken, der *Gene Ontology*, Literatur Kookkurrenzen und Strukturdaten kombiniert. Aufgrund der fortwährenden Informationsgewinnung in der Bioinformatik, ist anzunehmen, dass die Fülle an komplementäre Daten immer mehr zu nimmt, wodurch dieser Ansatz immer bessere und vollständigere Daten zum kombinieren bekommt.

Dies kann am besten erreicht werden, wenn die komplementären Daten direkt in die Optimierungsengine implementiert werden. Dabei ist es auch möglich aus dem statischen Schwellwert der GO-Terme einen dynamischen und genaueren Score zu für diesen komplementären Datentyp zu entwickeln. Zusätzlich kann dem Mangel an Interaktionsdaten durch die Implementierung von weiteren Interaktionsdatenbanken entgegen gewirkt werden.

Abschließend ist mit derart genaueren Daten auch eine Optimierung der aktivierungsabhänigen Daten (H/M Ansatz) möglich.

Literaturverzeichnis

Albert & Albert, 2004 — Albert I, Albert R: **Conserved network motifs allow protein-protein interaction prediction. Bioinformatics.** Bioinformatics. 2004 Dec 12;20(18):3346-52.

Alfarano et al., 2005 — Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, Hogue CW: **The Biomolecular Interaction Network Database and related tools 2005 update.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D418-24.

Aloy, 2007 — Aloy P: **Shaping the future of interactome networks.** Genome Biol 2007, 8(10):316.

Andreeva et al., 2007 — Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments.** Nucleic Acids Res. 2008 Jan;36(Database issue):D419-25.

Andreopoulos et al., 2007a — Andreopoulos B, An A, Wang X, Faloutsos M, Schroeder M: **Clustering by common friends finds locally significant proteins mediating modules.** Bioinformatics. 2007 May 1;23(9):1124-31.

Andreopoulos et al., 2007b — Andreopoulos B, An A, Wang X: **Hierarchical density-based clustering of categorical data and a simplification.** In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Springer LNCS 4426:11-22. Nanjing, China, May 22-25, 2007

Andreopoulos et al., 2009 — Andreopoulos B, Winter C, Labudde D, Schroeder M: **Triangle network motifs predict complexes by complementing high-error interactomes with structural information.** BMC Bioinformatics. 2009 Jun 27;10:196.

Ashburner et al., 2000 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** Nat Genet. 2000 May;25(1):25-9.

Barrett et al., 2005 Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D562-6.

Braun et al., 2009 Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M: **An experimentally derived confidence score for binary protein-protein interactions.** Nature Methods 6, 91 - 97 (2009)

Breitkreutz et al., 2003 Breitkreutz BJ, Stark C, Tyers M: **The GRID: the General Repository for Interaction Datasets.** Genome Biol. 2003;4(3):R23.

Boer et al., 2003 de Boer E, Rodriguez P, Bonte E, Krijgsveld J, Katsantoni E, Heck A, Grosveld F, Strouboulis J. **Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice.** Proc Natl Acad Sci U S A. 2003 Jun 24;100(13):7480-5.

Boulon et al., 2010 Boulon S, Ahmad Y, Trinkle-Mulcahy L, Verheggen C, Cobley A, Gregor P, Bertrand E, Whitehorn M, Lamond AI: **Establishment of a protein frequency library and its application in the reliable identification of specific protein interaction partners.** Mol Cell Proteomics. 2010 May;9(5):861-79.

Boutet et al., 2007 Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot.** Methods Mol Biol. 2007;406:89-112.

Brown & Jurisica, 2005 Brown KR, Jurisica I: **Online predicted human interaction database. Bioinformatics.** 2005 May 1;21(9):2076-82.

Brown & Jurisica, 2007 Brown KR, Jurisica I: **Unequal evolutionary conservation of human protein interactions in interologous networks.** Genome Biol. 2007;8(5):R95.

Carbon et al., 2009 Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO Hub; Web Presence Working Group. AmiGO: online access to ontology and annotation data.** Bioinformatics. 2009 Jan 15;25(2):288-9.

Charbonnier et al., 2008 Charbonnier S, Gallego O, Gavin AC: **The social network of a cell: recent advances in interactome mapping.** Biotechnol Annu Rev. 2008;14:1-28.

Chiang et al, 2007 Chiang T, Scholtens D, Sarkar D, Gentleman R, Huber W: **Coverage and error models of protein-protein interaction data by directed graph analysis.** Genome Biol 2007, 8(9):R186.

Chua et al., 2007 Chua HN, Sung WK, Wong L: **Using indirect protein interactions for the prediction of Gene Ontology functions.** BMC Bioinformatics. 2007 May 22;8 Suppl 4:S8.

Chua et al., 2008 Chua HN, Ning K, Sung WK, Leong HW, Wong L: **Using indirect protein-protein interactions for protein complex prediction.** J Bioinform Comput Biol. 2008 Jun;6(3):435-66.

Camacho et al., 2008 Camacho C, Madden T, Coulouris G, Ma N, Tao T, Agarwala R, Morgulis A: **BLAST Command Line Applications User Manual.** Created: June 23, 2008; Last Update: September 4, 2012. Bookshelf ID: NBK1763. Verfügbar von: http://www.ncbi.nlm.nih.gov/books/NBK1762/

Cormen et al.; 2007 Cormen TH, Leiserson C, Rivest RL, Stein C: **Algorithmen – Eine Einführung.** 2. Auflage. Oldenbourg Wissenschaftsverlag, München 2007, ISBN 978-3-486-58262-8, Kapitel 24 und 25.

Cote et al., 2006 Cote RG, Jones P, Apweiler R, Hermjakob H. **The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries.** BMC Bioinformatics. 2006 Feb 28;7(1):97 PMID: 16507094

Dabir et al., 2009 Dabir S, Kluge A, Dowlati A: (2009) **The association and nuclear translocation of the PIAS3-STAT3 complex is ligand and time dependent.** Mol Cancer Res. 2009 Nov;7(11):1854-60.

Dafas et al., 2004 Dafas P, Bolser D, Gomoluch J, Park J, Schroeder M: **Using convex hulls to extract interaction interfaces from known structures.** Bioinformatics, 20, 1486–1490.

Darwin, 1859 Darwin C: **On the Origin of Species by Means of natural Selection.** Murray London 1859

Deane et al., 2002 Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations.** Mol Cell Proteomics 2002, 1(5):349-56.

Degtyarenko et al., 2008 Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest.** Nucleic Acids Res. 2008 Jan;36(Database issue):D344-50.

de Matos et al, 2009 de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical entities of biological interest: an update.** Nucleic Acids Res. 2010 Jan;38(Database issue):D249-54.

D'haeseleer & Church, 2004 D'haeseleer P, Church GM: **Estimating and improving protein interaction error rates.** Proc IEEE Comput Syst_Bioinform Conf 2004:216-23.

Doms & Schroeder, 2005 Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology.** Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W783-6.

Du et al., 2009 Du Z, Li L, Chen CF, Yu PS, Wang JZ: **G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery.** Nucleic Acids Res. 2009 Jul;37(Web Server issue):W345-9.

Fearon et al., 1992 Fearon ER, Finkel T, Gillison ML, Kennedy SP, Casella JF, Tomaselli GF, Morrow JS, Van Dang C: **Karyoplasmic interaction selection strategy: a general strategy to detect protein-protein interactions in mammalian cells.** Proc Natl Acad Sci U S A. 1992 Sep 1;89(17):7958-62.

Flicek et al., 2012 Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM: **Ensembl 2012.** Nucleic Acids Res. 2012 Jan;40(Database issue):D84-90.

Finn et al., 2006 Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D247-51.

Geiger et al., 2011 Geiger T, Wisniewski JR, Cox J, Zanivan S, Kruger M, Ishihama Y, Mann M: **Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics.** Nat Protoc. 2011 Feb;6(2):147-57.

Gene Ontology Consortium, 2011 Gene Ontology Consortium: **The Gene Ontology: enhancements for 2011.** Nucleic Acids Res. 2012 Jan; 40(Database issue):D559-64.

Gong et al., 2005 Gong S, Yoon G, Insoo J, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K, Lee S, Choi H, Oh D, Lappe M, Holm L, Kim S, Bhak J: **PSIbase: the Database of the protein structural interactome MAP.** Bioinformatics. 2005 May 15;21(10):2541-3.

Gordon et al., 1998 Gordon GW, Berry G, Liang XH, Levine B, Herman B: **Quantitative fluorescence resonance energy transfer measurements using fluorescence microscopy.** Biophys J. 1998 May;74(5):2702-13.

Hart et al., 2006 Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** Genome_Biol 2006, 7(11):120.

Hoffmann & Valencia, 2003 Hoffmann R, Valencia A: **Protein interaction: same network, different hubs.** Trends Genet 2003, 19(12):681-3.

Holland, 1992 Holland JH: **Genetic algorithms**. Scientific American 1992, pp66-72

Holland et al., 2007 Holland SM, DeLeo FR, Elloumi HZ, Hsu AP, Uzel G, Brodsky N, Freeman AF, Demidowich A, Davis J, Turner ML, Anderson VL, Darnell DN, Welch PA, Kuhns DB, Frucht DM, Malech HL, Gallin JI, Kobayashi SD, Whitney AR, Voyich JM, Musser JM, Woellner C, Schäffer AA, Puck JM, Grimbacher B: **STAT3 mutations in the hyper-IgE syndrome.** N Engl J Med. 2007 Oct 18;357(16):1608-19.

Holmberg et al., 2005 Holmberg A, Blomstergren A, Nord O, Lukacs M, Lundeberg J, Uhlén M: **The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures.** Electrophoresis. 2005 Feb;26(3):501-10.

Hou et al., 2007 Hou T, Ray S, Brasier AR: **The functional role of an interleukin 6-inducible CDK9.STAT3 complex in human gamma-fibrinogen gene expression.** J Biol Chem. 2007 Dec 21;282(51):37091-102.

Howard et al., 1985 Howard PK, Shaw J, Otsuka AJ: **Nucleotide sequence of the birA gene encoding the biotin operon repressor and biotin holoenzyme synthetase functions of Escherichia coli.** Gene. 1985;35(3):321-31.

Jefferson et al., 2007 Jefferson ER, Walsh TP, Roberts TJ, Barton GJ: **SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions.** Nucleic Acids Res. 2007 Jan;35(Database issue):D580-9.

Jin et al., 2007 Jin G, Zhang S, Zhang XS, Chen L: **Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast.** PLoS One. 2007 Nov 21;2(11):e1207.

Li et al., 2006 Li H, Li J, Wong L: **Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale.** Bioinformatics. 2006 Apr 15;22(8):989-96

Licata et al., 2011 Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G: **MINT, the molecular interaction database:2012 update.** Nucleic Acids Res. 2012 Jan;40(Database issue):D857-61.

Kerrien et al., 2007 Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B1, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H: **Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions.** BMC Biol. 2007 Oct 9;5:44.

Kerrien et al., 2012 Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H: **The IntAct molecular interaction database in 2012.** Nucleic Acids Res. 2012 January; 40(D1): D841–D846.

Krüger et al., 2008 Krüger M, Moser M, Ussar S, Thievessen I, Luber CA, Forner F, Schmidt S, Zanivan S, Fässler R, Mann M: **SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function.** Cell. 2008 Jul 25;134(2):353-64.

Nissen, 1997 Nissen V: **Einführung in Evolutionäre Algorithmen.** Braunschweig: Viehweg, 1997

Phizicky & Fields, 1995 Phizicky EM, Fields S: **Protein-protein interactions: methods for detection and analysis.** Microbiol Rev. 1995 Mar;59(1):94-123.

Ranish et al.,2007 Ranish JA, Brand M, Aebersold R: **Using stable isotope tagging and mass spectrometry to characterize protein complexes and to detect changes in their composition.** Methods Mol Biol. 2007;359:17-35.

Rechenberg, 1973 Rechenberg I: **Evolutionsstrategie.** Friedrich Frommann Verlag Stuttgart, 1973

Rigaut et al., 1999 Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B: **A generic protein purification method for protein complex characterization and proteome exploration.** Nat Biotechnol. 1999 Oct;17(10):1030-2.

Rödel et al., 2000 Rödel B, Tavassoli K, Karsunky H, Schmidt T, Bachmann M, Schaper F, Heinrich P, Shuai K, Elsässer HP, Möröy T. **The zinc finger protein Gfi-1 can enhance STAT3 signaling by interacting with the STAT3 inhibitor PIAS3.** EMBO J. 2000 Nov 1;19(21):5845-55.

Goldberg, 1989 Goldberg DE: **Genetic Algorithms in Search, Optimization and Machine Learing.** Addison-Wesley Longman, Bonn, 1989

Gough et al., 2009 Gough DJ, Corlett A, Schlessinger K, Wegrzyn J, Larner AC, Levy DE: **Mitochondrial STAT3 supports Ras-dependent oncogenic transformation.** Science. 2009 Jun 26;324(5935):1713-6.

Haspel & Darnell, 1999 Haspel RL, Darnell JE Jr: **A nuclear protein tyrosine phosphatase is required for the inactivation of Stat1.** Proc Natl Acad Sci U S A. 1999 Aug 31;96(18):10188-93.

Hou et al., 2007 Hou T, Ray S, Brasier AR: **The functional role of an interleukin 6-inducible CDK9.STAT3 complex in human gamma-fibrinogen gene expression.** J Biol Chem. 2007 Dec 21;282(51):37091-102.

Kim et al., 2004 Kim WK, Bolser DM, Park JH: **Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP).** Bioinformatics, 20, 1138–1150.

McDowall et al., 2009 McDowall MD, Scott MS, Barton GJ: **PIPs: human protein-protein interaction prediction database.** Nucleic Acids Res. 2009 Jan;37(Database issue):D651-6.

Mertens et al., 2006 Mertens C, Zhong M, Krishnaraj R, Zou W, Chen X, Darnell JE Jr: **Dephosphorylation of phosphotyrosine on STAT1 dimers requires extensive spatial reorientation of the monomers facilitated by the N-terminal domain.** Genes Dev. 2006 Dec 15;20(24):3372-81.

Milo et al., 2002 Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** Science. 2002 Oct 25;298(5594):824-7.

Minegishi et al., 2007 Minegishi Y, Saito M, Tsuchiya S, Tsuge I, Takada H, Hara T, Kawamura N, Ariga T, Pasic S, Stojkovic O, Metin A, Karasuyama H. **Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome.** Nature. 2007 Aug 30;448(7157):1058-62.

Mishra et al., 2006 Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A: **Human protein reference database--2006 update.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D411-4.

Mulder et al., 2005 Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH: **InterPro, progress and status in 2005.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D201-5.

Oberbeck & Fogleman, 1989 Oberbeck VR, Fogleman G: **Estimates of the maximum time required to originate life.** Orig Life Evol Biosph. 1989;19(6):549-60.

O'Brien et al., 2005 — O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D476-80.

Okada et al., 2005 — Okada K, Kanaya S, Asai K: **Accurate extraction of functional associations between proteins based on common interaction partners and common domains.** Bioinformatics. 2005 May 1;21(9):2043-8.

Ong et al., 2002 — Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** Mol Cell Proteomics. 2002 May;1(5):376-86.

Orchard et al., 2012 — Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H: Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012 Apr;9(4):345-50.

Peng et al., 2006 — Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder.** BMC Bioinformatics. 2006 Apr 17;7:208.

Proietti et al., 2011 — Proietti CJ, Béguelin W, Flaqué MC, Cayrol F, Rivas MA, Tkach M, Charreau EH, Schillaci R, Elizalde PV: **Novel role of signal transducer and activator of transcription 3 as a progesterone receptor coactivator in breast cancer.** Steroids. 2011 Mar;76(4):381-92.

Ranish et al., 2007 — Ranish JA, Brand M, Aebersold R: **Using stable isotope tagging and mass spectrometry to characterize protein complexes and to detect changes in their composition.** Methods Mol Biol. 2007;359:17-35.

Rual et al., 2005 — Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network.** Nature. 2005 Oct;437(7062):1173-8.

Safran et al., 2010 Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A and Lancet D: **GeneCards Version 3: the human gene integrator.** Database (Oxford). 2010 Aug 5;2010:baq020.

Salwinski et al., 2004 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** Nucleic Acids Res. 2004 Jan 1;32(Database issue):D449-51.

Schatz, 1993 Schatz PJ: **Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in Escherichia coli.** Biotechnology (N Y). 1993 Oct;11(10):1138-43.

Schindler et al., 2007 Schindler C, Levy DE, Decker T: **JAK-STAT signaling: from interferons to cytokines.** J Biol Chem. 2007 Jul 13;282(28):20059-63.

Schöneburg, 1994 Schöneburg E: **Genetische Algorithmen und Evolutionsstrategien.** Bonn; Paris; Reading Mass: Addison-Wesley 1994

Scott et al., 2004 Scott MS, Thomas DY, Hallett MT: **Predicting subcellular localization via protein motif co-occurrence.** Genome Res. 2004 Oct;14(10A):1957-66.

Scott & Barton, 2007 Scott MS, Barton GJ: **Probabilistic prediction and ranking of human protein-protein interactions.** BMC Bioinformatics. 2007 Jul 5;8:239.

Shklar et al., 2005 Shklar M, Strichman-Almashanu L, Shmueli O, Shmoish M, Safran M, and Lancet D: **GeneTide--Terra Incognita Discovery Endeavor: a new transcriptome focused member of the GeneCards/GeneNote suite of databases.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D556-61.

Smith et al., 2007 Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** Nat Biotechnol. 2007 Nov;25(11):1251-5.

Sprinzak et al., 2003 Sprinzak E, Sattath S, Margalit H: **How reliable are experimental protein-protein interaction data?** J Mol Biol. 2003 Apr 11;327(5):919-23.

Stepkowski et al., 2008 — Stepkowski SM, Chen W, Ross JA, Nagy ZS, Kirken RA: **STAT3: an important regulator of multiple cytokine functions.** Transplantation. 2008 May 27;85(10):1372-7.

Stelzl et al., 2005 — Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker E: **A human protein-protein interaction network: a resource for annotating the proteome.** Cell 2005, 122(6):957-68.

Stelzl & Wanker, 2006 — Stelzl U, Wanker EE: **The value of high quality protein-protein interaction networks for systems biology.** Curr Opin Chem Biol. 2006 Dec;10(6):551-8.

Stumpf et al., 2008 — Stumpf MPH, Thorne T, Silva EdS, Stewart R, An HJ, Lappe M, Wiuf C**: Estimating the size of the human interactome.** Proc Natl Acad Sci USA 2008, 105(19):6959-64.

Su et al., 2004 — Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** Proc Natl Acad Sci U S A. 2004 Apr 20;101(16):6062-7.

Takezaki et al., 2012 — Takezaki S, Yamada M, Kato M, Park MJ, Maruyama K, Yamazaki Y, Chida N, Ohara O, Kobayashi I, Ariga T: **Chronic mucocutaneous candidiasis caused by a gain-of-function mutation in the STAT1 DNA-binding domain.** J Immunol. 2012 Aug 1;189(3):1521-6.

Trinkle-Mulcahy et al., 2008 — Trinkle-Mulcahy L, Boulon S, Lam YW, Urcia R, Boisvert FM, Vandermoere F, Morrice NA, Swift S, Rothbauer U, Leonhardt H, Lamond A: **Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes.** J Cell Biol. 2008 Oct 20;183(2):223-39.

Tsai et al., 1996 — Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **A dataset of protein–protein interfaces generated with a sequence-orderindependent comparison technique.** J. Mol. Biol., 260, 604–620.

Tsien, 1998 — Tsien RY: **The green fluorescent protein.** Annu Rev Biochem. 1998;67:509-44.

UniProt Consortium, 2012 — UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** Nucleic Acids Res. 2012 Jan;40(Database issue):D71-5.

Vandamme et al., 2011 Vandamme J, Völkel P, Rosnoblet C, Le Faou P, Angrand PO: **Interaction proteomics analysis of polycomb proteins defines distinct PRC1 complexes in mammalian cells.** Mol Cell Proteomics. 2011 Apr;10(4):M110.002642.

Verma et at., 2009 Verma NK, Dourlat J, Davies AM, Long A, Liu WQ, Garbay C, Kelleher D, Volkov Y: **STAT3-stathmin interactions control microtubule dynamics in migrating T-cells.** J Biol Chem. 2009 May 1;284(18):12349-62.

Vermeulen et al., 2006 Vermeulen M, Walter W, Le Guezennec X, Kim J, Edayathumangalam RS, Lasonder E, Luger K, Roeder RG, Logie C, Berger SL, Stunnenberg HG: **A feed-forward repression mechanism anchors the Sin3/histone deacetylase and N-CoR/SMRT corepressors on chromatin.** Mol Cell Biol. 2006 Jul;26(14):5226-36.

Vermeulen et al., 2008 Vermeulen M, Hubner NC, Mann M: **High confidence determination of specific protein-protein interactions using quantitative mass spectrometry.** Curr Opin Biotechnol. 2008 Aug;19(4):331-7.

Veverka et al., 2012 Veverka V, Baker T, Redpath NT, Carrington B, Muskett FW, Taylor RJ, Lawson AD, Henry AJ, Carr MD: **Conservation of functional sites on interleukin-6 and implications for evolution of signalling complex assembly and therapeutic intervention.** J Biol Chem. 2012 Oct 1. (published as Manuscript)

von Mering et al., 2003 von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** Nucleic Acids Res. 2003 Jan 1;31(1):258-61.

Wegenka et al., 1994 Wegenka UM, Lütticken C, Buschmann J, Yuan J, Lottspeich F, Müller-Esterl W, Schindler C, Roeb E, Heinrich PC, Horn F: **The interleukin-6-activated acute-phase response factor is antigenically and functionally related to members of the signal transducer and activator of transcription (STAT) family.** Mol Cell Biol.14(5) : 3186 – 96

Winter et al., 2006 Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D310-4.

Wu et al., 2006 Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: T**he Universal Protein Resource (UniProt): an expanding universe of protein information.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D187-91.

Xu et al., 1999 Xu Y, Piston DW, Johnson CH: **A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins.** Proc Natl Acad Sci U S A. 1999 Jan 5;96(1):151-6.

Yang et al, 2005 Yang SF, Yuan SS, Yeh YT, Wu MT, Su JH, Hung SC, Chai CY: **The role of p-STAT3 (ser727) revealed by its association with Ki-67 in cervical intraepithelial neoplasia.** Gynecol Oncol. 2005 Sep;98(3):446-52.

Yip & Gerstein, 2009 Yip KY, Gerstein M: **Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions.** Bioinformatics 2009, 25(2):243-50.

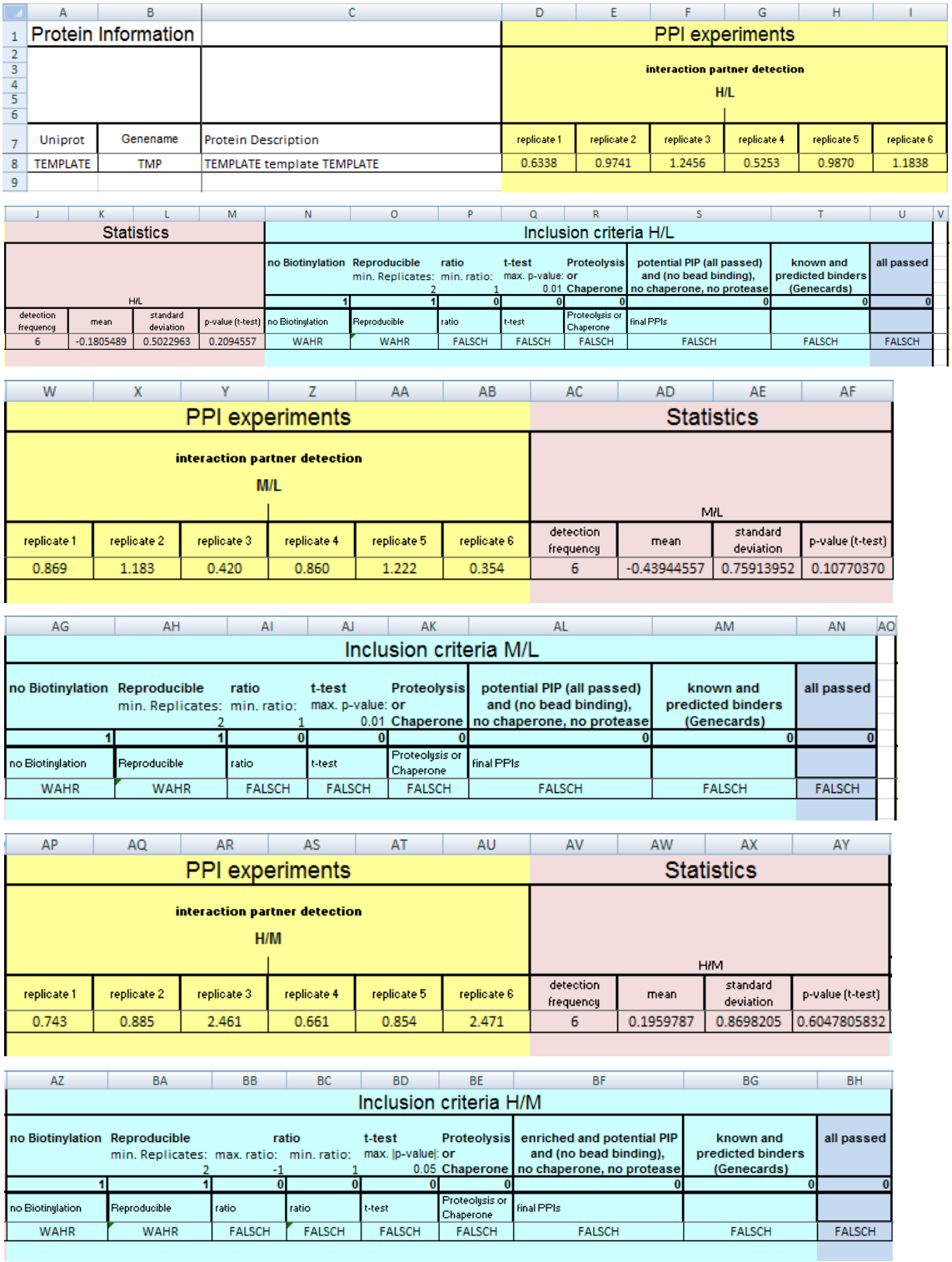
Yu & Cortez, 2011 Yu DS, Cortez D: **A role for CDK9-cyclin K in maintaining genome integrity.** Cell Cycle. 2011 Jan 1;10(1):28-32.

Zhang et al., 2006 Zhang S, Ning X, Zhang XS: **Identification of functional modules in a PPI network by clique percolation clustering.** Comput Biol Chem. 2006 Dec;30(6):445-51.

Zhang et al., 2008 Zhang Y, Xuan J, los Reyes BGdlR, Clarke R, Ressom HW: **Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data.** BMC Bioinformatics 2008, 9:203.

Anlagen

Anhang A: Excel-Template



Protein Information			PPI experiments					
			interaction partner detection H/L					
Uniprot	Genename	Protein Description	replicate 1	replicate 2	replicate 3	replicate 4	replicate 5	replicate 6
TEMPLATE	TMP	TEMPLATE template TEMPLATE	0.6338	0.9741	1.2456	0.5253	0.9870	1.1838

Statistics				Inclusion criteria H/L								
H/L				no Biotinylation	Reproducible min. Replicates: 2	ratio min. ratio: 1	t-test max. p-value: 0.01	Proteolysis or Chaperone	potential PIP (all passed) and (no bead binding), no chaperone, no protease	known and predicted binders (Genecards)	all passed
				1	1	0	0	0	0	0	0
detection frequency	mean	standard deviation	p-value (t-test)	no Biotinylation	Reproducible	ratio	t-test	Proteolysis or Chaperone	final PPIs		
6	-0.1805489	0.5022963	0.2094557	WAHR	WAHR	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH

PPI experiments						Statistics			
interaction partner detection M/L						M/L			
replicate 1	replicate 2	replicate 3	replicate 4	replicate 5	replicate 6	detection frequency	mean	standard deviation	p-value (t-test)
0.869	1.183	0.420	0.860	1.222	0.354	6	-0.43944557	0.75913952	0.10770370

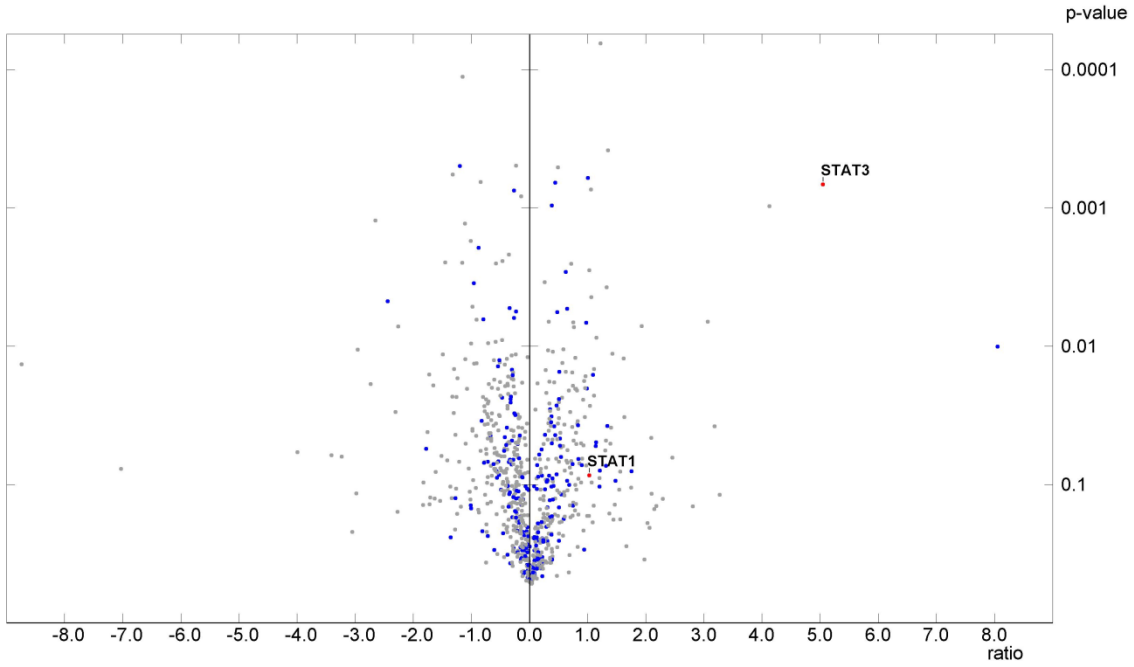
Inclusion criteria M/L							
no Biotinylation	Reproducible min. Replicates: 2	ratio min. ratio: 1	t-test max. p-value: 0.01	Proteolysis or Chaperone	potential PIP (all passed) and (no bead binding), no chaperone, no protease	known and predicted binders (Genecards)	all passed
1	1	0	0	0	0	0	0
no Biotinylation	Reproducible	ratio	t-test	Proteolysis or Chaperone	final PPIs		
WAHR	WAHR	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH

PPI experiments						Statistics			
interaction partner detection H/M						H/M			
replicate 1	replicate 2	replicate 3	replicate 4	replicate 5	replicate 6	detection frequency	mean	standard deviation	p-value (t-test)
0.743	0.885	2.461	0.661	0.854	2.471	6	0.1959787	0.8698205	0.6047805832

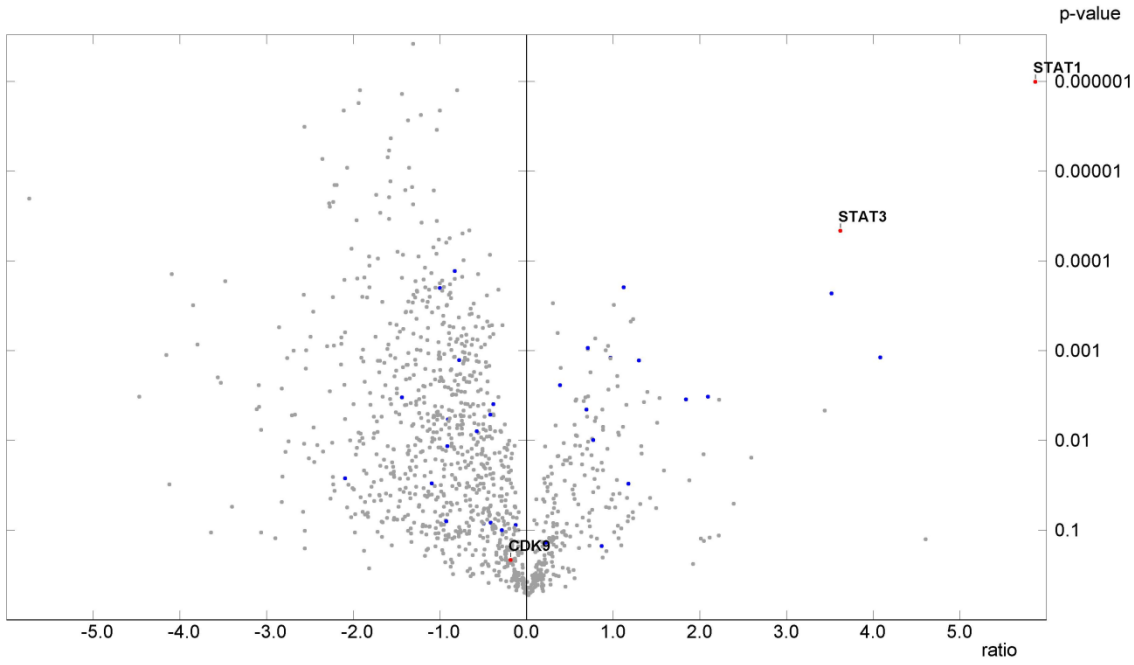
Inclusion criteria H/M								
no Biotinylation	Reproducible min. Replicates: 2	ratio max. ratio: -1	ratio min. ratio: 1	t-test max. |p-value|: 0.05	Proteolysis or Chaperone	enriched and potential PIP and (no bead binding), no chaperone, no protease	known and predicted binders (Genecards)	all passed
1	1	0	0	0	0	0	0	0
no Biotinylation	Reproducible	ratio	ratio	t-test	Proteolysis or Chaperone	final PPIs		
WAHR	WAHR	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH	FALSCH

Anhang B: Verteilung der GeneCards Einträge

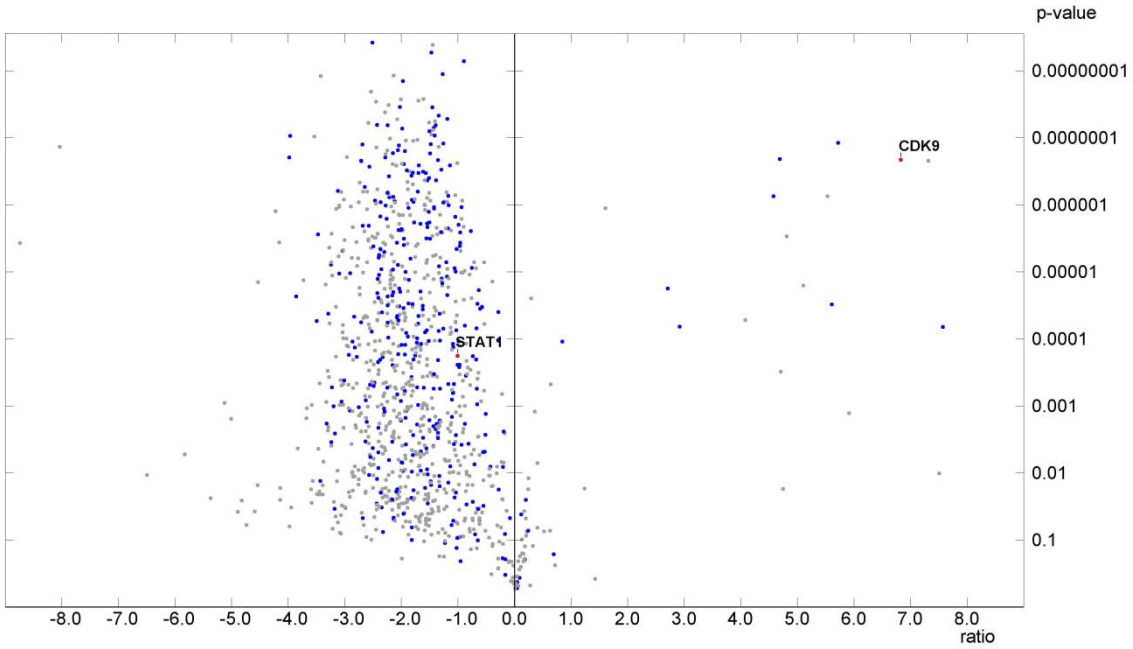
Diese Vulcano Plots zeigen von jedem Protein den H/L Ansatz, wobei STAT3, STAT1, CDK9 und BMI1 Rot und alle Einträge der GeneCards Blau markiert sind.



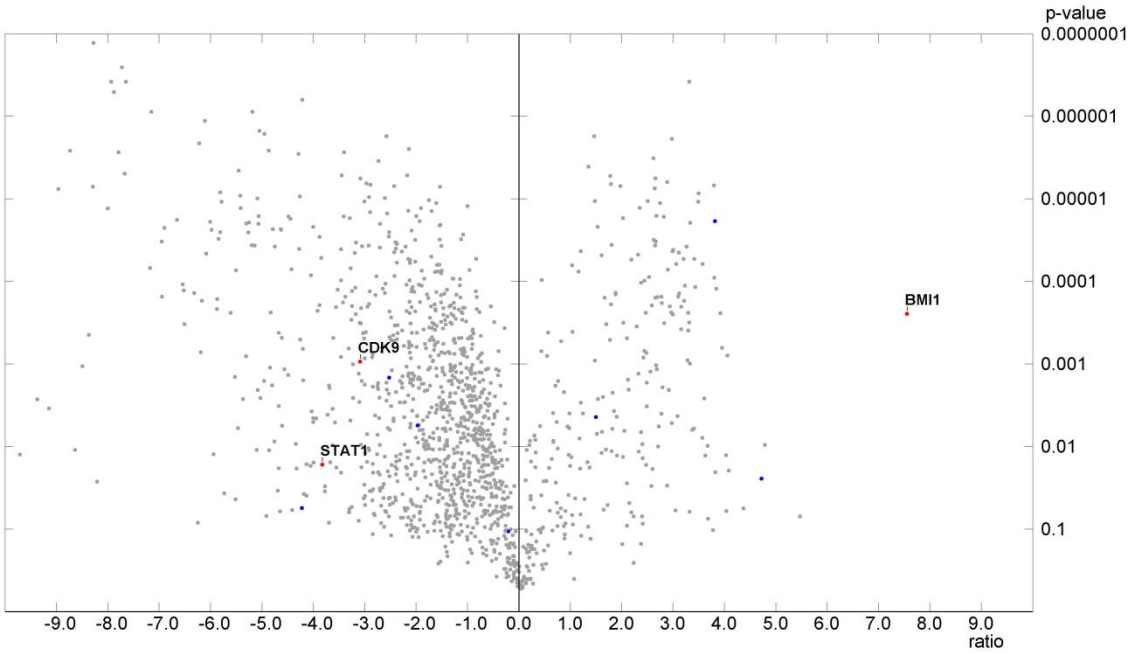
Vulcano Plot zu STAT3, H/L



Vulcano Plot zu STAT1, H/L



Vulcano Plot zu CDK9, H/L



Vulcano Plot zu BMI1, H/L

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 31. Oktober 2012 Riccardo Brumm